

---

# **Idcpy**

***Release 0.0.0***

**Alex Pinard, Allison Baker, Anderson Banihirwe, Dorit Hammerlin**

**Jan 11, 2022**



**CONTENTS:**

- 1 Reference to ldcpy paper 3**
  - 1.1 Installation using Conda (recommended) . . . . . 3
  - 1.2 Alternative Installation . . . . . 3
  - 1.3 Accessing the tutorial . . . . . 4
  - 1.4 Re-create notebooks with Pangeo Binder . . . . . 4
- 2 Documentation 5**
  - 2.1 Installation . . . . . 5
  - 2.2 Development . . . . . 6
  - 2.3 API Reference . . . . . 8
  - 2.4 Examples . . . . . 17
- 3 Indices and tables 91**
- Python Module Index 93**
- Index 95**





ldcpy is a utility for gathering and plotting metrics from NetCDF or Zarr files using the Pangeo stack. It also contains a number of statistical and visual tools for gathering metrics and comparing Earth System Model data files.

**AUTHORS** Alex Pinard, Allison Baker, Anderson Banihirwe, Dorit Hammerling

**COPYRIGHT** 2020 University Corporation for Atmospheric Research

**LICENSE** Apache 2.0

Documentation and usage examples are available [here](#).



## REFERENCE TO LDCPY PAPER

A. Pinard, D. M. Hammerling, and A. H. Baker. Assessing differences in large spatiotemporal climate datasets with a new Python package. In The 2020 IEEE International Workshop on Big Data Reduction, 2020. doi: 10.1109/BigData50022.2020.9378100.

Link to paper: <https://doi.org/10.1109/BigData50022.2020.9378100>

### 1.1 Installation using Conda (recommended)

Ensure conda is up to date and create a clean Python (3.6+) environment:

```
conda update conda
conda create --name ldcpy python=3.8
conda activate ldcpy
```

Now install ldcpy:

```
conda install -c conda-forge ldcpy
```

### 1.2 Alternative Installation

Ensure pip is up to date, and your version of python is at least 3.6:

```
pip install --upgrade pip
python --version
```

Install cartopy using the instructions provided at <https://scitools.org.uk/cartopy/docs/latest/installing.html>.

Then install ldcpy:

```
pip install ldcpy
```

## 1.3 Accessing the tutorial

If you want access to the tutorial notebook, clone the repository (this will create a local repository in the current directory):

```
git clone https://github.com/NCAR/ldcpy.git
```

Start by enabling Hinterland for code completion and code hinting in Jupyter Notebook and then opening the tutorial notebook:

```
jupyter nbextension enable hinterland/hinterland
jupyter notebook
```

The tutorial notebook can be found in docs/source/notebooks/TutorialNotebook.ipynb, feel free to gather your own metrics or create your own plots in this notebook!

Other example notebooks that use the sample data in this repository include PopData.ipynb and MetricsNotebook.ipynb.

The AWSDataNotebook grabs data from AWS, so can be run on a laptop with the caveat that the files are large.

The following notebooks assume that you are using NCAR's JupyterHub (<https://jupyterhub.ucar.edu>): Large-DataGladernotebook.ipynb, CompressionSamples.ipynb, and error\_bias.ipynb

## 1.4 Re-create notebooks with Pangeo Binder

Try the notebooks hosted in this repo on Pangeo Binder. Note that the session is ephemeral. Your home directory will not persist, so remember to download your notebooks if you make changes that you need to use at a later time!

Note: All example notebooks are in docs/source/notebooks (the easiest ones to use in binder first are TutorialNotebook.ipynb and PopData.ipynb)

## DOCUMENTATION

## 2.1 Installation

### 2.1.1 Installation using Conda (recommended)

Ensure conda is up to date and create a clean Python (3.6+) environment:

```
conda update conda
conda create --name ldcpy python=3.8
conda activate ldcpy
```

Now install ldcpy:

```
conda install -c conda-forge ldcpy
```

### 2.1.2 Alternative Installation

Ensure pip is up to date, and your version of python is at least 3.6:

```
pip install --upgrade pip
python --version
```

Install cartopy using the instructions provided at <https://scitools.org.uk/cartopy/docs/latest/installing.html>.

Then install ldcpy:

```
pip install ldcpy
```

### 2.1.3 Accessing the tutorial (for users)

If you want access to the tutorial notebook, clone the repository (this will create a local repository in the current directory):

```
git clone https://github.com/NCAR/ldcpy.git
```

Start by activating the ldcpy environment, enabling Hinterland for code completion in Jupyter Notebook and then starting the notebook server:

```
conda activate ldcpy
conda install -c conda-forge jupyter_contrib_nbextensions
jupyter contrib nbextension install --user
jupyter nbextension enable hinterland/hinterland
jupyter notebook
```

The tutorial notebook can be found in docs/source/notebooks/TutorialNotebook.ipynb, feel free to gather your own metrics or create your own plots in this notebook!

## 2.2 Development

### 2.2.1 Installation for Developers

First, clone the repository and cd into the root of the repository:

```
git clone https://github.com/NCAR/ldcpy.git
cd ldcpy
```

For a development install, do the following in the ldcpy repository directory:

```
conda env update -f environment.yml
conda activate ldcpy
python -m pip install -e .
```

Then install the pre-commit script and git hooks for code style checking:

```
pre-commit install
```

This code block enables optional extensions for code completion, code hinting and minimizing tracebacks in Jupyter. Then start the jupyter notebook server in your browser (at localhost:8888):

```
jupyter nbextension enable hinterland/hinterland
jupyter nbextension enable skip-traceback/main

conda activate ldcpy
jupyter notebook
```

### 2.2.2 Instructions and Tips for Contributing

For viewing changes to documentation in the repo, do the following:

```
pip install -r docs/requirements.txt
sphinx-reload docs/
```

This starts and opens a local version of the documentation in your browser (at localhost:5500/index.html) and keeps it up to date with any changes made. Note that changes to docstrings in the code will not trigger an update, only changes to the .rst files in the docs/ folder.

If you have added a feature or fixed a bug, add new tests to the appropriate file in the tests/ directory to test that the feature is working or that the bug is fixed. Before committing changes to the code, run all tests from the project root directory to ensure they are passing.

```
pytest -n 4
```

Additionally, rerun the TutorialNotebook in Jupyter (Kernel -> Restart & Run All). Check that no unexpected behavior is encountered in these plots.

Now you are ready to commit your code. pre-commit should automatically run black, flake8, and isort to enforce style guidelines. If changes are made, the first commit will fail and you will need to stage the changes that have been made before committing again. If, for some reason, pre-commit fails to make changes to your files, you should be able to run the following to clean the files manually:

```
black --skip-string-normalization --line-length=100 .
flake8 .
isort .
```

### 2.2.3 Adding new package dependencies to ldcpy

- 1) Adding new package dependencies requires updating the code in the following four places:

/ci/environment.yml /ci/environment-dev.yml /ci/upstream-dev-environment.yml /requirements.txt

If the package dependency is specifically used for documentation, instead of adding it to /requirements.txt, add it to:

/docs/source/requirements.txt

If this package is only used for documentation, skip the remaining steps.

- 2) If the package is one that includes C code (such as numpy or scipy), update the autodoc\_mock\_imports list in /docs/source/conf.py. The latest build of the documentation can be found at (<https://readthedocs.org/projects/ldcpy/builds/>), if the build fails and the error message indicates a problem with the newest package - try adding it to autodoc\_mock\_imports.
- 3) Finally, update the ldcpy-feedstock repository (git clone <https://github.com/conda-forge/ldcpy-feedstock.git>), or manually create a branch and add the dependency in the browser. Name the branch add-<new\_dependency\_name>. In the file /recipe/meta.yaml, in the “requirements” section, under “run”, add your dependency to the list.
- 4) If the CI build encounters errors after adding a dependency, check the status of the CI workflow at (<https://github.com/NCAR/ldcpy/actions?query=workflow%3ACI>) to determine if the error is related to the new package.

### 2.2.4 Creating a Release

Updating the package on PyPi:

- 1) On the ldcpy Github page, select Releases on the right sidebar, and select “Draft a new release”
- 2) Create a new tag by incrementing the minor or major version number. Give the release a title and description.
- 3) Publish the release. Check the Actions tab -> Upload Package to PyPi workflow to ensure it completes.

Updating the package on Conda Forge:

- 0) Ensure the package has been updated on PyPi.
- 1) Fork the ldcpy\_feedstock repository (<https://github.com/conda-forge/ldcpy-feedstock>)
- 2) In recipe/meta.yaml, set the version number to match the latest release tag. Make sure the build number is 0 if you are changing the version number.
- 3) In recipe/meta.yaml, update the sha256 hash. The hash for the latest release can be found at <https://pypi.org/project/ldcpy/#files>. Copy the hash from ldcpy-x.xx.xx.tar.gz.

- 4) In recipe/meta.yml, add any new package dependencies under the run section of the requirements.
- 5) From your fork's github page, create a pull request pointed at the conda\_forge repository.
- 6) Make sure each step listed in the pull request checklist is completed. See [https://conda-forge.org/docs/maintainer/updating\\_pkgs.html](https://conda-forge.org/docs/maintainer/updating_pkgs.html) if needed.
- 7) Allow some time for all the tests to complete, these take between 8-20 minutes. See the error/warning output if any tests fail.
- 8) Merge the pull request. The new version will be available on conda forge shortly.

## 2.3 API Reference

This page provides an auto-generated summary of ldcpy's API. For more details and examples, refer to the relevant chapters in the main part of the documentation.

### 2.3.1 ldcpy Util (ldcpy.util)

`ldcpy.util.check_metrics(ds, varname, set1, set2, ks_tol=0.05, pcc_tol=0.99999, spre_tol=5.0, ssim_tol=0.995, **calcs_kwargs)`

Check the K-S, Pearson Correlation, and Spatial Relative Error calcs

#### Parameters

- **ds** (*xarray.Dataset*) – An xarray dataset containing multiple netCDF files concatenated across a 'collection' dimension
- **varname** (*str*) – The variable of interest in the dataset
- **set1** (*str*) – The collection label of the "control" data
- **set2** (*str*) – The collection label of the (1st) data to compare
- **ks\_tol** (*float, optional*) – The p-value threshold (significance level) for the K-S test (default = .05)
- **pcc\_tol** (*float, optional*) – The default Pearson correlation coefficient (default = .99999)
- **spre\_tol** (*float, optional*) – The percentage threshold for failing grid points in the spatial relative error test (default = 5.0).
- **ssim\_tol** (*float, optional*) – The threshold for the data ssim test (default = .995)
- **\*\*calcs\_kwargs** – Additional keyword arguments passed through to the Datasetcalcs instance.

#### Returns out

**Return type** Number of failing calcs



## Notes

Check the K-S, Pearson Correlation, and Spatial Relative Error calcs from:

A. H. Baker, H. Xu, D. M. Hammerling, S. Li, and J. Clyne, “Toward a Multi-method Approach: Lossy Data Compression for Climate Simulation Data”, in J.M. Kunkel et al. (Eds.): ISC High Performance Workshops 2017, Lecture Notes in Computer Science 10524, pp. 30–42, 2017 (doi:10.1007/978-3-319-67630-2\_3).

Check the Data SSIM, which is a modification of SSIM calc from:

A.H. Baker, D.M. Hammerling, and T.L. Turton. “Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data”, Computer Graphics Forum 38(3), June 2019, pp. 517-528 (doi:10.1111/cgf.13707).

K-S: fail if p-value < .05 (significance level) Pearson correlation coefficient: fail if coefficient < .99999 Spatial relative error: fail if > 5% of grid points fail relative error Data SSIM: fail if Data SSIM < .995

`ldcpy.util.collect_datasets(data_type, varnames, list_of_ds, labels, **kwargs)`

Concatenate several different xarray datasets across a new “collection” dimension, which can be accessed with the specified labels. Stores them in an xarray dataset which can be passed to the ldcpy plot functions (Call this OR `open_datasets()` before plotting.)

### Parameters

- **data\_type** (*string*) – Current data types: :cam-fv, pop
- **varnames** (*list*) – The variable(s) of interest to combine across input files (usually just one)
- **list\_of\_datasets** (*list*) – The datasets to be concatenated into a collection
- **labels** (*list*) – The respective label to access data from each dataset (also used in plotting fcn)
- **\*\*kwargs** : (optional) – Additional arguments passed on to `xarray.concat()`. A list of available arguments can be found here: <https://xarray-test.readthedocs.io/en/latest/generated/xarray.concat.html>

**Returns** **out** – a collection containing all the data from the list datasets

**Return type** `xarray.Dataset`

`ldcpy.util.compare_stats(ds, varname: str, sets, significant_digits: int = 5, include_ssim: bool = False, weighted: bool = True, **calcs_kwargs)`

Print error summary statistics for multiple DataArrays (should just be a single time slice)

### Parameters

- **ds** (`xarray.Dataset`) – An xarray dataset containing multiple netCDF files concatenated across a ‘collection’ dimension
- **varname** (*str*) – The variable of interest in the dataset
- **sets** (*list of str*) – The labels of the collection to compare (all will be compared to the first set)
- **significant\_digits** (*int, optional*) – The number of significant digits to use when printing stats (default 5)
- **include\_ssim** (*bool, optional*) – Whether or not to compute the image ssim - slow for 3D vars (default: False)
- **weighted** (*bool, optional*) – Whether or not weight the means (default = True)

- **\*\*calcs\_kwargs** – Additional keyword arguments passed through to the Datasetcalcs instance.

**Returns out**

**Return type** `None`

`ldcpy.util.open_datasets(data_type, varnames, list_of_files, labels, weights=True, **kwargs)`

Open several different netCDF files, concatenate across a new ‘collection’ dimension, which can be accessed with the specified labels. Stores them in an xarray dataset which can be passed to the ldcpy plot functions.

**Parameters**

- **data\_type** (*string*) – Current data types: :cam-fv, pop
- **varnames** (*list*) – The variable(s) of interest to combine across input files (usually just one)
- **list\_of\_files** (*list*) – The file paths for the netCDF file(s) to be opened
- **labels** (*list*) – The respective label to access data from each netCDF file (also used in plotting fcns)
- **\*\*kwargs** – (optional) – Additional arguments passed on to `xarray.open_mfdataset()`. A list of available arguments can be found here: [http://xarray.pydata.org/en/stable/generated/xarray.open\\_dataset.html](http://xarray.pydata.org/en/stable/generated/xarray.open_dataset.html)

**Returns out** – a collection containing all the data from the list of files

**Return type** `xarray.Dataset`

`ldcpy.util.save_metrics(full_ds, varname, set1, set2, time=0, color='coolwarm', lev=0, ks_tol=0.05, pcc_tol=0.99999, spre_tol=5.0, ssim_tol=0.9995, location='names.csv')`

Check the K-S, Pearson Correlation, and Spatial Relative Error metrics from: A. H. Baker, H. Xu, D. M. Hammerling, S. Li, and J. Clyne, “Toward a Multi-method Approach: Lossy Data Compression for Climate Simulation Data”, in J.M. Kunkel et al. (Eds.): ISC High Performance Workshops 2017, Lecture Notes in Computer Science 10524, pp. 30–42, 2017 (doi:10.1007/978-3-319-67630-2\_3). Check the SSIM metric from: A.H. Baker, D.M. Hammerling, and T.L. Turton. “Evaluating image quality measures to assess the impact of lossy data compression applied to climate simulation data”, Computer Graphics Forum 38(3), June 2019, pp. 517-528 (doi:10.1111/cgf.13707). Default tolerances for the tests are: ————— K-S: fail if p-value < .05 (significance level) Pearson correlation coefficient: fail if coefficient < .99999 Spatial relative error: fail if > 5% of grid points fail relative error SSIM: fail if SSIM < .99995 :param ds: An xarray dataset containing multiple netCDF files concatenated across a ‘collection’ dimension :type ds: xarray.Dataset :param varname: The variable of interest in the dataset :type varname: str :param set1: The collection label of the “control” data :type set1: str :param set2: The collection label of the (1st) data to compare :type set2: str :param time: The time index used t (default = 0) :type time: int, optional :param ks\_tol: The p-value threshold (significance level) for the K-S test (default = .05) :type ks\_tol: float, optional :param pcc\_tol: The default Pearson correlation coefficient (default = .99999) :type pcc\_tol: float, optional :param spre\_tol: The percentage threshold for failing grid points in the spatial relative error test (default = 5.0). :type spre\_tol: float, optional :param ssim\_tol: The threshold for the ssim test (default = .99995) :type ssim\_tol: float, optional :param time: The level index of interest in a 3D dataset (default 0) :type time: lev, optional

**Returns out**

**Return type** Number of failing metrics

`ldcpy.util.subset_data(ds, subset=None, lat=None, lon=None, lev=None, start=None, end=None, time_dim_name='time', vertical_dim_name=None, lat_coord_name=None, lon_coord_name=None)`

Get a subset of the given dataArray, returns a dataArray

### 2.3.2 ldcpy Plot (ldcpy.plot)

```
class ldcpy.plot.calcsPlot(ds, varname, calc, sets, group_by=None, scale='linear', calc_type='raw',
                           plot_type='spatial', transform='none', subset=None, approx_lat=None,
                           approx_lon=None, lev=0, color='coolwarm', standardized_err=False,
                           quantile=None, contour_levs=24, short_title=False, axes_symmetric=False,
                           legend_loc='upper right', vert_plot=False, tex_format=False,
                           legend_offset=None, weighted=True, basic_plot=False)
```

This class contains code to plot calcs in an xarray Dataset that has either 'lat' and 'lon' dimensions, or a 'time' dimension.

```
time_series_plot(da_sets, titles)
    time series plot
```

```
ldcpy.plot.plot(ds, varname, calc, sets, group_by=None, scale='linear', calc_type='raw', plot_type='spatial',
                transform='none', subset=None, lat=None, lon=None, lev=0, color='coolwarm',
                quantile=None, start=None, end=None, short_title=False, axes_symmetric=False,
                legend_loc='upper right', vert_plot=False, tex_format=False, legend_offset=None,
                weighted=True, basic_plot=False)
```

Plots the data given an xarray dataset

#### Parameters

- **ds** (*xarray.Dataset*) – The input dataset
- **varname** (*str*) – The name of the variable to be plotted
- **calc** (*str*) – The name of the calc to be plotted (must match a property name in the Datasetcalcs class in ldcpy.plot, for more information about the available calcs see ldcpy.Datasetcalcs) Acceptable values include:
  - ns\_con\_var
  - ew\_con\_var
  - mean
  - std
  - variance
  - prob\_positive
  - prob\_negative
  - odds\_positive
  - zscore
  - mean\_abs
  - mean\_squared
  - rms
  - sum
  - sum\_squared
  - corr\_lag1
  - quantile
  - lag1
  - standardized\_mean

- ann\_harmonic\_ratio
- pooled\_variance\_ratio
- **sets** (*list* <*str*>) – The labels of the dataset to gather calcs from
- **group\_by** (*str*) – how to group the data in time series plots. Valid groupings:
  - time.day
  - time.dayofyear
  - time.month
  - time.year
- **scale** (*str*, *optional*) – time-series y-axis plot transformation. (default “linear”) Valid options:
  - linear
  - log
- **calc\_type** (*str*, *optional*) – The type of operation to be performed on the calcs. (default ‘raw’) Valid options:
  - raw: the unaltered calc values
  - diff: the difference between the calc values in the first set and every other set
  - ratio: the ratio of the calc values in (2nd, 3rd, 4th... sets/1st set)
  - calc\_of\_diff: the calc value computed on the difference between the first set and every other set
- **plot\_type** (*str* , *optional*) – The type of plot to be created. (default ‘spatial’) Valid options:
  - spatial: a plot of the world with values at each lat and lon point (takes the mean across the time dimension)
  - time-series: A time-series plot of the data (computed by taking the mean across the lat and lon dimensions)
  - histogram: A histogram of the time-series data
- **transform** (*str*, *optional*) – data transformation. (default ‘none’) Valid options:
  - none
  - log
- **subset** (*str*, *optional*) – subset of the data to gather calcs on (default None). Valid options:
  - first5: the first 5 days of data
  - DJF: data from the months December, January, February
  - MAM: data from the months March, April, May
  - JJA: data from the months June, July, August
  - SON: data from the months September, October, November
- **lat** (*float* , *optional*) – The latitude of the data to gather calcs on (default None).
- **lon** (*float* , *optional*) – The longitude of the data to gather calcs on (default None).

- **lev** (*float*, *optional*) – The level of the data to gather calcs on (used if plotting from a 3d data set), (default 0).
- **color** (*str*, *optional*) – The color scheme for spatial plots, (default ‘coolwarm’). see [https://matplotlib.org/3.1.1/gallery/color/colormap\\_reference.html](https://matplotlib.org/3.1.1/gallery/color/colormap_reference.html) for more options
- **quantile** (*float*, *optional*) – A value between 0 and 1 required if calc=”quantile”, corresponding to the desired quantile to gather, (default 0.5).
- **start** (*int*, *optional*) – A value between 0 and the number of time slices indicating the start time of a subset, (default None).
- **end** (*int*, *optional*) – A value between 0 and the number of time slices indicating the end time of a subset, (default None)
- **calc\_ssim** (*bool*, *optional*) – Whether or not to calculate the ssim (structural similarity index) between two plots (only applies to plot\_type = ‘spatial’), (default False)
- **short\_title** (*bool*, *optional*) – If True, use a shortened title in the plot output (default False).
- **axes\_symmetric** (*bool*, *optional*) – Whether or not to make the colorbar axes symmetric about zero (used in a spatial plot) (default False)
- **legend\_loc** (*str*, *optional*) – The location to put the legend in a time-series plot in single-column format (plot\_type = “time\_series”, vert\_plot=True) (default “upper right”)
- **vert\_plot** (*bool*, *optional*) – If true, forces plots into a single column format and enlarges text. (default False)
- **tex\_format** (*bool*, *optional*) – Whether to interpret all plot output strings as latex formatting (default False)
- **legend\_offset** (*2-tuple*, *optional*) – The x- and y- offset of the legend. Moves the corner of the legend specified by legend\_loc to the specified location specified (where (0,0) is the bottom left corner of the plot and (1,1) is the top right corner). Only affects time-series, histogram, and periodogram plots.

**Returns out**

**Return type** `None`

`ldcpy.plot.tex_escape(text)`

**Parameters** `text` – a plain text message

**Returns** the message escaped to appear correctly in LaTeX

### 2.3.3 Idcpy Metrics (`ldcpy.metrics`)

```
class ldcpy.calcs.Datasetcalcs(ds: xarray.core.dataarray.DataArray, data_type: str, aggregate_dims: list,
                               time_dim_name: str = 'time', lat_dim_name: Optional[str] = None,
                               lon_dim_name: Optional[str] = None, vert_dim_name: Optional[str] =
                               None, lat_coord_name: Optional[str] = None, lon_coord_name:
                               Optional[str] = None, q: float = 0.5, spre_tol: float = 0.0001,
                               weighted=True)
```

This class contains calcs for each point of a dataset after aggregating across one or more dimensions, and a method to access these calcs. Expects a DataArray.

```
get_calc(name: str, q: Optional[int] = 0.5, grouping: Optional[str] = None, ddof=1)
```

Gets a calc aggregated across one or more dimensions of the dataset

**Parameters**

- **name** (*str*) – The name of the calc (must be identical to a property name)
- **q** (*float*, *optional*) – (default 0.5)

**Returns out** – A DataArray of the same size and dimensions the original dataarray, minus those dimensions that were aggregated across.

**Return type** `xarray.DataArray`

**get\_single\_calc**(*name: str*)

Gets a calc consisting of a single float value

**Parameters name** (*str*) – the name of the calc (must be identical to a property name)

**Returns out** – The calc value

**Return type** `float`

**property annual\_harmonic\_relative\_ratio: xarray.core.dataarray.DataArray**

The annual harmonic relative to the average periodogram value in a neighborhood of 50 frequencies around the annual frequency NOTE: This assumes the values along the “time” dimension are equally spaced. NOTE: This calc returns a lat-lon array regardless of aggregate dimensions, so can only be used in a spatial plot.

**property annual\_harmonic\_relative\_ratio\_pct\_sig: numpy.ndarray**

The percentage of points past the significance cutoff (p value  $\leq 0.01$ ) for the annual harmonic relative to the average periodogram value in a neighborhood of 50 frequencies around the annual frequency

**property cdf: xarray.core.dataarray.DataArray**

The empirical CDF of the dataset.

**property entropy: xarray.core.dataarray.DataArray**

An estimate for the entropy of the data (using gzip) # lower is better (1.0 means random - no compression possible)

**property ew\_con\_var: xarray.core.dataarray.DataArray**

The East-West Contrast Variance averaged along the aggregate dimensions

**property lag1: xarray.core.dataarray.DataArray**

The deseasonalized lag-1 autocorrelation value by day of year NOTE: This calc returns an array of spatial values as the data set regardless of aggregate dimensions, so can only be plotted in a spatial plot.

**property lag1\_first\_difference: xarray.core.dataarray.DataArray**

The deseasonalized lag-1 autocorrelation value of the first difference of the data by day of year NOTE: This calc returns an array of spatial values as the data set regardless of aggregate dimensions, so can only be plotted in a spatial plot.

**property lat\_autocorr: xarray.core.dataarray.DataArray**

the correlation of a variable with itself shifted in the latitude dimension

**Type** Autocorrelation

**property lev\_autocorr: xarray.core.dataarray.DataArray**

the correlation of a variable with itself shifted in the vertical dimension

**Type** Autocorrelation

**property lon\_autocorr: xarray.core.dataarray.DataArray**

the correlation of a variable with itself shifted in the longitude dimension

**Type** Autocorrelation

**property mae\_day\_max: xarray.core.dataarray.DataArray**  
 The day of maximum mean absolute value at the point. NOTE: only available in spatial and spatial comparison plots

**property mean: xarray.core.dataarray.DataArray**  
 The mean along the aggregate dimensions

**property mean\_abs: xarray.core.dataarray.DataArray**  
 The mean of the absolute errors along the aggregate dimensions

**property mean\_squared: xarray.core.dataarray.DataArray**  
 The absolute value of the mean along the aggregate dimensions

**property most\_repeated: xarray.core.dataarray.DataArray**  
 Most repeated value in dataset

**property most\_repeated\_percent: xarray.core.dataarray.DataArray**  
 Most repeated value in dataset

**property n\_s\_first\_differences: xarray.core.dataarray.DataArray**  
 First differences along the west-east direction

**property ns\_con\_var: xarray.core.dataarray.DataArray**  
 The North-South Contrast Variance averaged along the aggregate dimensions

**property num\_negative: xarray.core.dataarray.DataArray**  
 The probability that a point is negative

**property num\_positive: xarray.core.dataarray.DataArray**  
 The probability that a point is positive

**property num\_zero: xarray.core.dataarray.DataArray**  
 The probability that a point is zero

**property odds\_positive: xarray.core.dataarray.DataArray**  
 The odds that a point is positive =  $\text{prob\_positive} / (1 - \text{prob\_positive})$

**property percent\_unique: xarray.core.dataarray.DataArray**  
 Percentage of unique values in the dataset

**property pooled\_variance: xarray.core.dataarray.DataArray**  
 The overall variance of the dataset

**property pooled\_variance\_ratio: xarray.core.dataarray.DataArray**  
 The pooled variance along the aggregate dimensions

**property prob\_negative: xarray.core.dataarray.DataArray**  
 The probability that a point is negative

**property prob\_positive: xarray.core.dataarray.DataArray**  
 The probability that a point is positive

**property range: xarray.core.dataarray.DataArray**  
 The range of the dataset

**property root\_mean\_squared: xarray.core.dataarray.DataArray**  
 The absolute value of the mean along the aggregate dimensions

**property standardized\_mean: xarray.core.dataarray.DataArray**  
 The mean at each point along the aggregate dimensions divided by the standard deviation NOTE: will always be 0 if aggregating over all dimensions

**property std: xarray.core.dataarray.DataArray**  
 The standard deviation along the aggregate dimensions

**property variance: xarray.core.dataarray.DataArray**

The variance along the aggregate dimensions

**property w\_e\_derivative: xarray.core.dataarray.DataArray**

Derivative of dataset from west-east

**property w\_e\_first\_differences: xarray.core.dataarray.DataArray**

First differences along the west-east direction

**property zscore: xarray.core.dataarray.DataArray**

The z-score of a point averaged along the aggregate dimensions under the null hypothesis that the true mean is zero. NOTE: currently assumes we are aggregating along the time dimension so is only suitable for a spatial plot.

**property zscore\_cutoff: numpy.ndarray**

The Z-Score cutoff for a point to be considered significant

**property zscore\_percent\_significant: numpy.ndarray**

The percent of points where the zscore is considered significant

**class** ldcpy.calcs.Diffcalcs(*ds1: xarray.core.dataarray.DataArray, ds2: xarray.core.dataarray.DataArray, data\_type: str, aggregate\_dims: Optional[list] = None, \*\*calcs\_kwargs*)

This class contains calcs on the overall dataset that require more than one input dataset to compute

**get\_diff\_calc**(*name: str, color: Optional[str] = 'coolwarm'*)

Gets a calc on the dataset that requires more than one input dataset

**Parameters** *name* (*str*) – The name of the calc (must be identical to a property name)

**Returns** *out*

**Return type** *float*

**property covariance: xarray.core.dataarray.DataArray**

The covariance between the two datasets

**property ks\_p\_value**

The Kolmogorov-Smirnov p-value

**property max\_spatial\_rel\_error**

We compute the relative error at each grid point and return the maximum.

**property normalized\_max\_pointwise\_error**

The absolute value of the maximum pointwise difference, normalized by the range of values for the first set

**property normalized\_root\_mean\_squared**

The absolute value of the mean along the aggregate dimensions, normalized by the range of values for the first set

**property pearson\_correlation\_coefficient**

returns the pearson correlation coefficient between the two datasets

**property spatial\_rel\_error**

At each grid point, we compute the relative error. Then we report the percentage of grid point whose relative error is above the specified tolerance (1e-4 by default).

**property ssim\_value**

We compute the SSIM (structural similarity index) on the visualization of the spatial data. This creates two plots and uses the standard SSIM.

**property ssim\_value\_fp\_fast**

Faster implementation than `ssim_value_fp_orig` (this is the default DSSIM option).



**property ssim\_value\_fp\_orig**

To mimic what zchecker does - the ssim on the fp data with original constants and no scaling (so-called “straightforward” approach. This will return Nan on POP data or CAM data with NaNs because scikit SSIM fuction does not handle NaNs.

**property ssim\_value\_fp\_slow**

We compute the SSIM (structural similarity index) on the spatial data - using the data itself (we do not create an image) - this is the slower non-matrix implementation that is good for experiementing (not in practice).

## 2.4 Examples

### 2.4.1 Tutorial

ldcpy is a utility for gathering and plotting derived quantities from NetCDF or Zarr files using the Pangeo stack. This tutorial notebook targets comparing CESM data in its original form to CESM data that has undergone lossy compression (meaning that the reconstructed file is not exactly equivalent to the original file). The tools provided in ldcpy are intended to highlight differences due to compression artifacts in order to assist scientist in evaluating the amount of lossy compression to apply to their data.

The CESM data used here are NetCDF files in “timeseries” file format, meaning that each NetCDF file contains one (major) output variable (e.g., surface temperature or precipitation rate) and spans multiple timesteps (daily, monthly, 6-hourly, etc.). CESM timeseries files are regularly released in large public datasets.

```
[1]: # Add ldcpy root to system path
import sys

import astropy

sys.path.insert(0, '../..../')

# Import ldcpy package
# Autoreloads package everytime the package is called, so changes to code will be
↪reflected in the notebook if the above sys.path.insert(...) line is uncommented.
%load_ext autoreload
%autoreload 2

# suppress all of the divide by zero warnings
import warnings

warnings.filterwarnings("ignore")

import ldcpy

# display the plots in this notebook
%matplotlib inline
```

## Overview

This notebook demonstrates the use of ldcpy on the sample data included with this package. It explains how to open datasets (and view metadata), display basic statistics about the data, and create both time-series and spatial plots of the datasets and related quantities. Plot examples start out with the essential arguments, and subsequent examples explore the additional plotting options that are available.

For information about installation, see [these instructions](#), and for information about usage, see the API reference [here](#).

## Loading Datasets and Viewing Metadata

The first step in comparing the data is to load the data from the files that we are interested into a “collection” for ldcpy to use. To do this, we use `ldcpy.open_datasets()`. This function requires the following three arguments:

- *varnames* : the variable(s) of interest to combine across files (typically the timeseries file variable name)
- *list\_of\_files* : a list of full file paths (either relative or absolute)
- *labels* : a corresponding list of names (or labels) for each file in the collection

Note: This function is a wrapper for `xarray.open_mfdatasets()`, and any additional key/value pairs passed in as a dictionary are used as arguments to `xarray.open_mfdatasets()`. For example, specifying the chunk size (“chunks”) will be important for large data (see `LargeDataGladeNotebook.ipynb` for more information and an example).

We setup three different collections of timeseries datasets in these examples:

- *col\_ts* contains daily surface temperature (TS) data (2D data) for 100 days
- *col\_prect* contains daily precipitation rate (PRECT) data (2D data) for 60 days
- *col\_t* contains monthly temperature (T) data (3D data) for 3 months

These datasets are collections of variable data from several different netCDF files, which are given labels in the third parameter to the `ldcpy.open_datasets()` function. These names/labels can be whatever you want (e.g., “orig”, “control”, “bob”, ...), but they should be informative because the names will be used to select the appropriate dataset later and as part of the plot titles.

In this example, in each dataset collection we include a file with the original (uncompressed) data as well as additional file(s) with the same data subject to different levels of lossy compression.

*Note: If you do not need to get the data from files (e.g., you have already used `xarray.open_dataset()`), then use `ldcpy.collect_datasets()` instead of `ldcpy.open_datasets` (see example in `AWSDDataNotebook.ipynb`).*

```
[2]: # col_ts is a collection containing TS data
col_ts = ldcpy.open_datasets(
    "cam-fv",
    ["TS"],
    [
        "../.../data/cam-fv/orig.TS.100days.nc",
        "../.../data/cam-fv/zfp1.0.TS.100days.nc",
        "../.../data/cam-fv/zfp1e-1.TS.100days.nc",
    ],
    ["orig", "zfpA1.0", "zfpA1e-1"],
)
# col_prect contains PRECT data
col_prect = ldcpy.open_datasets(
    "cam-fv",
    ["PRECT"],
    [
```

(continues on next page)

(continued from previous page)

```

        "../.../data/cam-fv/orig.PRECT.60days.nc",
        "../.../data/cam-fv/zfp1e-7.PRECT.60days.nc",
        "../.../data/cam-fv/zfp1e-11.PRECT.60days.nc",
    ],
    ["orig", "zfpA1e-7", "zfpA1e-11"],
)
# col_t contains 3D T data (here we specify the chunk to be a single timeslice)
col_t = ldcpy.open_datasets(
    "cam-fv",
    ["T"],
    [
        "../.../data/cam-fv/cam-fv.T.3months.nc",
        "../.../data/cam-fv/c.fzip.cam-fv.T.3months.nc",
    ],
    ["orig", "comp"],
    chunks={"time": 1},
)

```

dataset size in GB 0.07

dataset size in GB 0.04

dataset size in GB 0.04

Note that running the `open_datasets` function (as above) prints out the size of each dataset collection. For `col_prect`, the `chunks` parameter is used by DASK (which is further explained in `LargeDataGladeNotebook.ipynb`).

Printing a dataset collection reveals the dimension names, sizes, datatypes and values, among other metadata. The dimensions and the length of each dimension are listed at the top of the output. Coordinates list the dimensions vertically, along with the data type of each dimension and the coordinate values of the dimension (for example, we can see that the 192 latitude data points are spaced evenly between -90 and 90). Data variables lists all the variables available in the dataset. For these timeseries files, only the one (major) variable will be of interest. For `col_t`, that variable is temperature (T), which was specified in the first argument of the `open_datasets()` call. The so-called major variable will have the required “lat”, “lon”, and “time” dimensions. If the variable is 3D (as in this example), a “lev” dimension will indicate that the dataset contains values at multiple altitudes (here, `lev=30`). Finally, a “collection” dimension indicates that we concatenated several datasets together. (In this `col_t` example, we concatenated 2 files together.)

```
[3]: # print information about col_t
col_t
```

```
[3]: <xarray.Dataset>
Dimensions:      (collection: 2, time: 3, lev: 30, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lev          (lev) float64 3.643 7.595 14.36 24.61 ... 957.5 976.3 992.6
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 1920-02-01 00:00:00 ... 1920-04-01 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>,
↳ meta=np.ndarray>
  * collection   (collection) <U4 'orig' 'comp'
Data variables:
```

(continues on next page)

(continued from previous page)

```

T          (collection, time, lev, lat, lon) float32 dask.array<chunksize=(1, 1, 30,
→ 192, 288), meta=np.ndarray>
Attributes: (12/14)
  Conventions:    CF-1.0
  source:         CAM
  case:           b.e11.B20TRC5CNBDRD.f09_g16.031
  title:          UNSET
  logname:        mickelso
  host:           ys0219
  ...
  initial_file:   b.e11.B20TRC5CNBDRD.f09_g16.001.cam.i.1920-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:        Thu Jul  9 14:15:11 2020: ncks -d time,0,2,1 cam-fv.T.6...
  NCO:            netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures:  area: cell_area
  data_type:      cam-fv

```

## Comparing Summary Statistics

The `compare_stats` function can be used to compute and compare the overall statistics for a **single** timeslice in two (or more) datasets from a collection. To use this function, three arguments are required. In order, they are:

- *ds* - a single time slice in a collection of datasets read in from `ldcpy.open_datasets()` or `ldcpy.collect_datasets()`
- *varname* - the variable name we want to get statistics for (in this case ‘TS’ is the variable in our dataset collection *col\_ts* )
- *sets* - the labels of the datasets in the collection we are interested in comparing (if more than two, all are compared to the first one)

Additionally, three optional arguments can be specified:

- *significant\_digits* - the number of significant digits to print (default = 5)
- *include\_ssim* - include the image ssim (default = False. Note this takes a bit of time for 3D vars)
- *weighted* - use weighted averages (default = True)

```

[10]: # print 'TS' statistics about 'orig', 'zfpA1.0', and 'zfpA1e-1 and diff between the 'orig' and
→ the other two datasets
# for time slice = 0
ds = col_ts.isel(time=0)
ldcpy.compare_stats(ds, "TS", ["orig", "zfpA1.0", "zfpA1e-1"], significant_digits=6)

```

<IPython.core.display.HTML object>

	orig	zfpA1.0	zfpA1e-1
mean	284.49	284.481	284.489
variance	534.015	533.69	533.995
standard deviation	23.1088	23.1017	23.1083
min value	216.741	216.816	216.747
max value	315.584	315.57	315.576
probability positive	1	1	1
number of zeros	0	0	0
spatial autocorr - latitude	0.993918	0.993911	0.993918

(continues on next page)

(continued from previous page)

spatial autocorr - longitude	0.996801	0.996791	0.996801
entropy estimate	0.414723	0.247491	0.347534
<IPython.core.display.HTML object>			
<IPython.core.display.HTML object>			
	zfpA1.0	zfpA1e-1	
max abs diff	0.405884	0.0229187	
min abs diff	0	0	
mean abs diff	0.0565128	0.0042215	
mean squared diff	7.32045e-05	3.04995e-07	
root mean squared diff	0.0729538	0.00532525	
normalized root mean squared diff	0.000761543	5.39821e-05	
normalized max pointwise error	0.00410636	0.00023187	
pearson correlation coefficient	0.999995	1	
ks p-value	1	1	
spatial relative error(% > 0.0001)	68.958	0	
max spatial relative error	0.00147397	8.11948e-05	
Data SSIM	0.981813	0.998227	

```
[ ]: # without weighted means
ldcpy.compare_stats(ds, "TS", ["orig", "zfpA1.0", "zfpA1e-1"], significant_digits=6,
↳weighted=False)
```

We can also generate derived quantities on a particular dataset. While this is done “behind the scenes” with the plotting functions, we first demonstrate here how the user can access this data without creating a plot.

We use an object of type `ldcpy.Datasetcalcs` to gather calculations derived from a dataset. To create an `ldcpy.DatasetMetrics` object, we first grab the particular dataset from our collection that we are interested in (in the usual xarray manner). For example, the following will grab the data for the TS variable labeled ‘orig’ in the `col_ts` dataset that we created:

```
[ ]: # get the orig dataset
my_data = col_ts["TS"].sel(collection="orig")
my_data.attrs["data_type"] = col_ts.data_type
my_data.attrs["set_name"] = "orig"
```

```
[20]: my_data
```

```
[20]: <xarray.DataArray 'TS' (time: 100, lat: 192, lon: 288)>
dask.array<getitem, shape=(100, 192, 288), dtype=float32, chunksize=(100, 192, 288),
↳chunktype=numpy.ndarray>
Coordinates:
  * lat      (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon      (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time     (time) object 1920-01-01 00:00:00 ... 1920-04-10 00:00:00
    cell_area (lat, lon) float64 dask.array<chunksize=(192, 288), meta=np.ndarray>
    collection <U8 'orig'
Attributes:
  units:      K
  long_name:  Surface temperature (radiative)
  cell_methods: time: mean
```

(continues on next page)

(continued from previous page)

```
data_type:    cam-fv
set_name:     orig
```

```
[23]: my_data.data_type
```

```
[23]: 'cam-fv'
```

Then we create a Datasetcalcs object using the data and a list of dimensions that we want to aggregate the data along. We usually want to aggregate all of the timeslices (“time”) or spatial points (“lat” and “lon”):

```
[ ]: ds_calcs = ldcpy.Datasetcalcs(my_data, [])
new_col_ts = ds_calcs.get_calc_ds("derivative", "TS_deriv")
```

```
dataset size in GB 0.04
```

```
[25]: new_col_ts
```

```
[25]: <xarray.Dataset>
Dimensions:    (collection: 1, lat: 192, lon: 288, time: 100)
Coordinates:
  * lat        (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon        (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time       (time) object 1920-01-01 00:00:00 ... 1920-04-10 00:00:00
    cell_area  (lon) object None None None None None ... None None None None
  * collection (collection) <U4 'orig'
Data variables:
    TS_deriv   (collection, time, lat, lon) float64 dask.array<chunksize=(1, 100, 192, ↵
↵288), meta=np.ndarray>
Attributes:
    cell_measures:  area: cell_area
    data_type:      cam-fv
```

```
[28]: ds_calcs_across_time = ldcpy.Datasetcalcs(my_data, "cam-fv", ["time"])
ds_calcs_across_space = ldcpy.Datasetcalcs(my_data, "cam-fv", ["lat", "lon"])
```

```
[46]: ds_calcs_pointwise
```

```
[46]: (<xarray.Dataset>
Dimensions:    (collection: 1, lat: 192, lon: 288, time: 100)
Coordinates:
  * lat        (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon        (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time       (time) object 1920-01-01 00:00:00 ... 1920-04-10 00:00:00
    cell_area  (lon) object None None None None None ... None None None None
  * collection (collection) <U4 'orig'
Data variables:
    m          (collection, time, lat, lon) float64 dask.array<chunksize=(1, 100, 192, ↵
↵288), meta=np.ndarray>
    dtype      object float64
Attributes:
    cell_measures:  area: cell_area
    data_type:      cam-fv,
[ ])
```

```
[ ]: d
```

Now when we call the `get_calc()` method on this class, a quantity will be computed across each of these specified dimensions. For example, below we compute the “mean” across time.

```
[32]: my_data_mean_across_time = ds_calcs_across_space.get_calc_ds("mean", "m")

# trigger computation
my_data_mean_across_time

-----
KeyError                                Traceback (most recent call last)
<ipython-input-32-0afa051f7c00> in <module>
----> 1 my_data_mean_across_time = ds_calcs_across_space.get_calc_ds("mean", "m")
      2
      3 # trigger computation
      4 my_data_mean_across_time

~/git/ldcpy/ldcpy/calcs.py in get_calc_ds(self, calc_name, var_name)
    917     da = self.get_calc(calc_name)
    918     ds = da.squeeze().to_dataset(name=var_name)
--> 919     new_ds = collect_datasets(self._ds.data_type, [var_name], [ds],
    920                             [self._ds.set_name])
    921

~/git/ldcpy/ldcpy/collect_datasets.py in collect_datasets(data_type, varnames, list_of_
-> ds, labels, **kwargs)
    64     full_ds.coords['cell_area'] = (
    65         xr.DataArray(full_ds.variables.mapping.get(weights_name))
--> 66         .expand_dims(lon=full_ds.dims['lon'])
    67         .transpose()
    68     )

/opt/miniconda3/envs/ldcpy_a/lib/python3.8/site-packages/xarray/core/utils.py in __
-> getitem__(self, key)
    424
    425     def __getitem__(self, key: K) -> V:
--> 426         return self.mapping[key]
    427
    428     def __iter__(self) -> Iterator[K]:

/opt/miniconda3/envs/ldcpy_a/lib/python3.8/site-packages/xarray/core/utils.py in __
-> getitem__(self, key)
    455
    456     def __getitem__(self, key: K) -> V:
--> 457         return self.mapping[key]
    458
    459     def __setitem__(self, key: K, value: V) -> None:

KeyError: 'lon'
```

```
[ ]: # Here just ask for the spatial mean at the first time step
my_data_mean_across_space = ds_calcs_across_space.get_calc("mean").isel(time=0)
```

(continues on next page)

(continued from previous page)

```
# trigger computation
my_data_mean_across_space.load()
```

```
[13]: # Here just ask for the spatial mean at the first time step
      d = ds_calcs_across_space.get_single_calc("most_repeated_percent")
      # trigger computation
      d
```

```
[13]: 4.159432870370371e-06
```

There are many currently available quantities to choose from. A complete list of derived quantities that can be passed to `get_calc()` is available [here](#).

## Spatial Plots

### A Basic Spatial Plot

First we demonstrate the most basic usage of the `ldcpy.plot()` function. In its simplest form, `ldcpy.plot()` plots a single spatial plot of a dataset (i.e., no comparison) and requires:

- *ds* - the collection of datasets read in from `ldcpy.open_datasets()` (or `ldcpy.collect_datasets()`)
- *varname* - the variable name we want to get statistics for
- *sets* - a list of the labels of the datasets in the collection that we are interested in
- *calc* - the name of the calculation to be plotted

There are a number of optional arguments as well, that will be demonstrated in subsequent plots. These options include:

- *group\_by*
- *scale*
- *calc\_type*
- *plot\_type*
- *transform*
- *subset*
- *lat*
- *lon*
- *lev*
- *color*
- *quantile*
- *start*
- *end*

A full explanation of each optional argument is available in the documentation [here](#) - as well as all available options. By default, a spatial plot of this data is created. The title of the plot contains the name of the dataset, the variable of interest, and the calculation that is plotted.



The following command generates a plot showing the mean TS (surface temperature) value from our dataset collection `col_ts` over time at each point in the dataset labeled 'orig'. Note that plots can be saved by using the matplotlib function `savefig`:

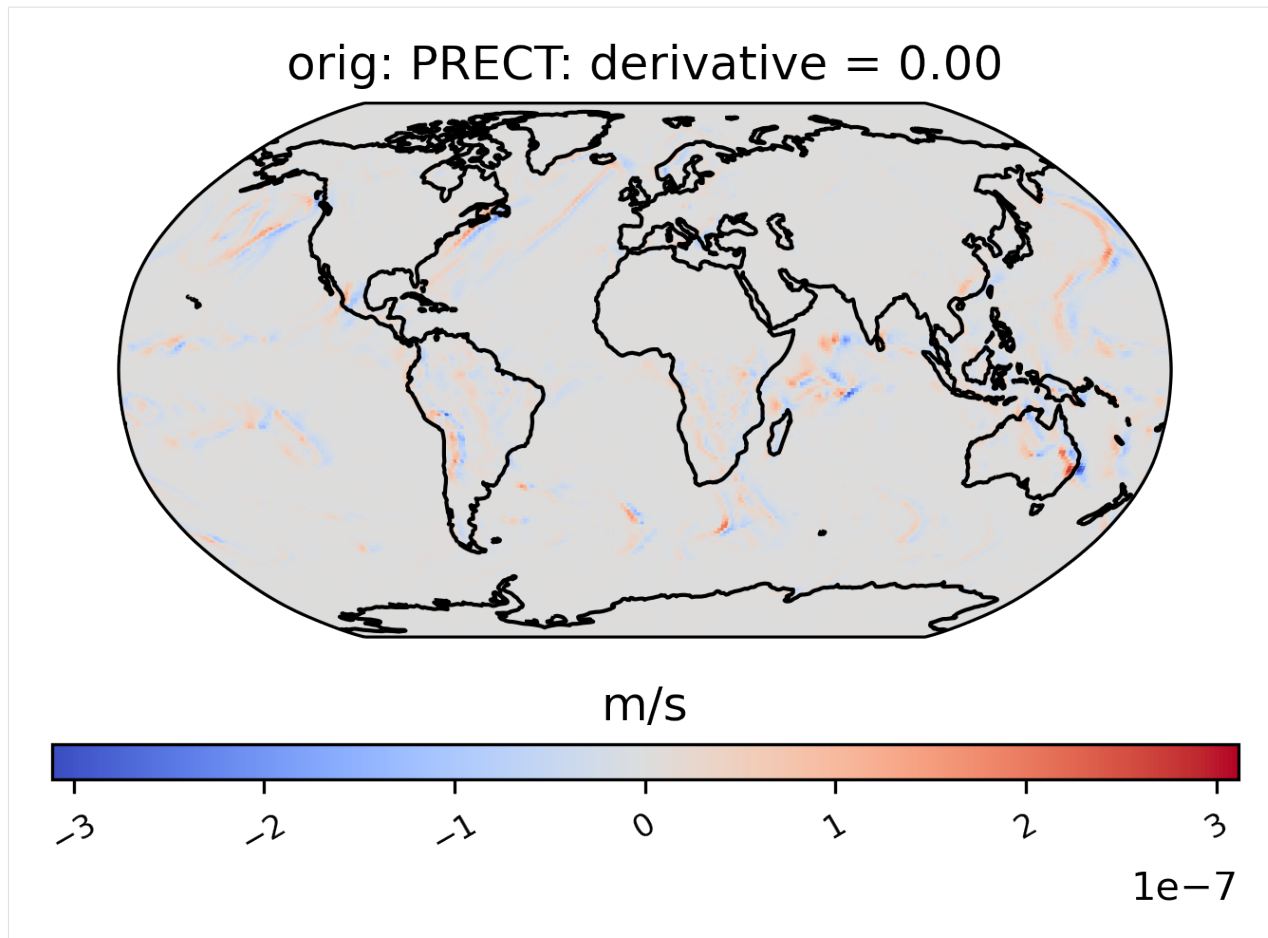
```
[ ]: sample = my_data.values.flatten()
     ecdf = sm.distributions.ECDF(sample)

     x = np.linspace(min(sample), max(sample))
     y = ecdf(x)
     plt.step(x, y)
     plt.show()
```

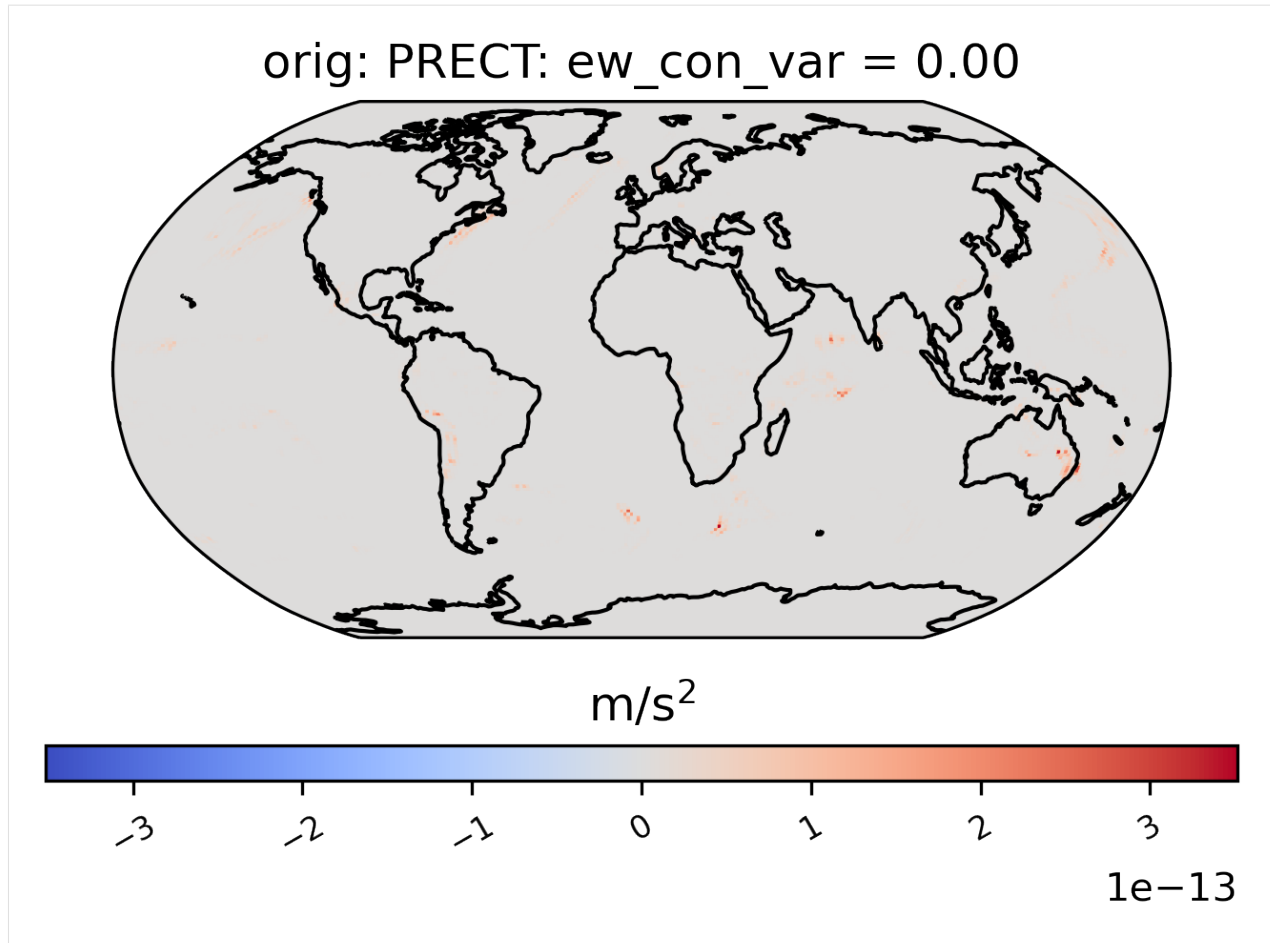
```
[25]: y
```

```
[25]: array([1.80844907e-07, 1.80844907e-07, 1.26591435e-06, 3.25520833e-06,
            1.37442130e-05, 2.03812211e-04, 8.34418403e-04, 2.38932292e-03,
            5.10832610e-03, 8.75759549e-03, 1.30620660e-02, 1.83053024e-02,
            2.55072700e-02, 3.47191479e-02, 4.66787833e-02, 6.20352286e-02,
            8.14572483e-02, 1.05130570e-01, 1.29461625e-01, 1.54272099e-01,
            1.77338505e-01, 1.96113824e-01, 2.13090097e-01, 2.28529550e-01,
            2.42921911e-01, 2.56257776e-01, 2.68729203e-01, 2.80904586e-01,
            2.92346101e-01, 3.03636068e-01, 3.16497758e-01, 3.35356807e-01,
            3.70174334e-01, 4.09271738e-01, 4.56041124e-01, 5.01259766e-01,
            5.41031359e-01, 5.73883102e-01, 6.07587167e-01, 6.46929073e-01,
            6.95648510e-01, 7.64410265e-01, 8.63390299e-01, 9.65394604e-01,
            9.94603769e-01, 9.97698387e-01, 9.99332321e-01, 9.99875579e-01,
            9.99992224e-01, 1.00000000e+00])
```

```
[84]: ldcpy.plot(
        col_prect,
        "PRECT",
        sets=["orig"],
        calc="derivative",
        plot_type="spatial",
        start=0,
        end=1,
        weighted=False,
        axes_symmetric=True,
    )
    # Uncomment to save this image to the filesystem
    # import matplotlib.pyplot as plt
    # plt.savefig(f"MYFIGURE.png", bbox_inches='tight')
```



```
[85]: ldcpy.plot(
    col_prect,
    "PRECT",
    sets=["orig"],
    calc="ew_con_var",
    plot_type="spatial",
    start=0,
    end=1,
    weighted=False,
    axes_symmetric=True,
)
# Uncomment to save this image to the filesystem
# import matplotlib.pyplot as plt
# plt.savefig(f"MYFIGURE.png", bbox_inches='tight')
```



We can also plot quantities other than the mean, such as the standard deviation at each grid point over time. We can also change the color scheme (for a full list of quantities and color schemes, see the [documentation](#)). Here is an example of a plot of the same dataset from `col_ts` as above but using a different color scheme and the standard deviation calculation. Notice that we do not specify the `'plot_type'` this time, because it defaults to `'spatial'`:

```
[ ]: # plot of the standard deviation of TS values in the col_ds "orig" dataset
ldcpy.plot(col_ts, "TS", sets=["orig"], calc="std", color="cmo.thermal")
```

## Spatial Subsetting

Plotting a derived quantity of a subset of the data (e.g., not all of the time slices in the dataset) is possible using the `subset` keyword. In the plot below, we just look at “DJF” data (Dec., Jan., Feb.). Other options for `subset` are available [here](#).

```
[ ]: # plot of mean winter TS values in the col_ts "orig" dataset
ldcpy.plot(col_ts, "TS", sets=["orig"], calc="mean", subset="DJF")
```

It is also possible to plot calculations for a subset of the time slices by specifying the start and end indices of the time we are interested in. This command creates a spatial plot of the mean TS values in “orig” for the first five days of data:

```
[ ]: # plot of the first 5 TS values in the ds orig dataset
ldcpy.plot(col_ts, "TS", sets=["orig"], calc="mean", start=0, end=5)
```

Finally, for a 3D dataset, we can specify which vertical level to view using the “lev” keyword. Note that “lev” is a dimension in our dataset *col\_t* (see printed output for *col\_t* above), and in this case lev=30, meaning that lev ranges from 0 and 29, where 0 is at the surface (by default, lev=0):

```
[ ]: # plot of T values at lev=29 in the col_t "orig" dataset
ldcpy.plot(col_t, "T", sets=["orig"], calc="variance", lev=29)
```

## Spatial Comparisons and Diff Plots

If we want a side-by-side comparison of two datasets, we need to specify an additional dataset name in the sets argument. The plot below shows the mean TS value over time at each point in the ‘orig’ (original), ‘zfpA1.0’ (compressed with zfp, absolute error tolerance 1.0) and ‘zfpA1e-1’ (compressed with zfp, absolute error tolerance 0.1) datasets in collection *col\_ts*. The “vert\_plot” argument indicates that the arrangement of the subplots should be vertical.

```
[ ]: # comparison between mean TS values in col_ts for "orig" and "zfpA1.0" datasets
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0", "zfpA1e-1"],
    calc="mean",
    vert_plot=True,
)
```

To the eye, these plots look identical. This is because the effects of compression are small compared to the magnitude of the data. We can view the compression artifacts more clearly by plotting the difference between two plots. This can be done by setting the calc\_type to ‘diff’:

```
[ ]: # diff between mean TS values in col_ts for "orig" and "zfpA1.0" datasets
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0", "zfpA1e-1"],
    calc="mean",
    calc_type="diff",
)
```

We are not limited to comparing side-by-side plots of the original and compressed data. It is also possible to compare two different compressed datasets side-by-side as well, by using a different dataset name for the first element in sets:

```
[ ]: # comparison between variance of TS values in col_ts for "zfpA1e-1" and "zfpA1.0" datasets
↪
ldcpy.plot(col_ts, "TS", sets=["zfpA1e-1", "zfpA1.0"], calc="variance")
```

```
[ ]: # diff between mean TS values in col_ts for "zfpA1e-1" and "zfpA1.0" datasets
ldcpy.plot(
    col_ts,
    "TS",
    sets=["zfpA1e-1", "zfpA1.0"],
    calc="mean",
    calc_type="diff",
)
```

Sometimes comparison plots can look strikingly different, indicating a potential problem with the compression. This plot shows the probability of negative rainfall (PRECT). We would expect this quantity to be zero everywhere on the globe (because negative rainfall does not make sense!), but the compressed output shows regions where the probability is significantly higher than zero:

```
[ ]: ldcpy.plot(
    col_prect,
    "PRECT",
    sets=["orig", "zfpA1e-11"],
    calc="prob_negative",
    color="binary",
)

[ ]: my_data = col_prect["PRECT"].sel(collection="orig")
ds_calcs_across_time = ldcpy.Datasetcalcs(my_data, ["time"])
prob_neg = ds_calcs_across_time.get_calc("prob_negative")

# trigger computation
prob_neg.load()
```

### Unusual Values (inf, -inf, NaN)

Some derived quantities result in values that are +/- infinity, or NaN (likely resulting from operations like 0/0 or inf/inf). NaN values are plotted in neon green, infinity is plotted in white, and negative infinity is plotted in black (regardless of color scheme). If infinite values are present in the plot data, arrows on either side of the colorbar are shown to indicate the color for +/- infinity. This plot shows the log of the ratio of the odds of positive rainfall over time in the compressed and original output,  $\log(\text{odds\_positive compressed} / \text{odds\_positive original})$ . Here we are suppressing all of the divide by zero warnings for aesthetic reasons.

The following plot showcases some interesting plot features. We can see sections of the map that take NaN values, and other sections that are black because taking the log transform has resulted in many points with the value -inf:

```
[ ]: # plot of the ratio of odds positive PRECT values in col_prect zfpA1.0 dataset / col_
↳prect orig dataset (log transform)
ldcpy.plot(
    col_prect,
    "PRECT",
    sets=["orig", "zfpA1e-11"],
    calc="odds_positive",
    calc_type="ratio",
    transform="log",
    axes_symmetric=True,
    vert_plot=True,
)
```

If all values are NaN, then the colorbar is not shown. Instead, a legend is shown indicating the green(!) color of NaN values, and the whole plot is colored gray. (If all values are infinite, then the plot is displayed normally with all values either black or white). Because the example dataset only contains 60 days of data, the deseasonalized lag-1 values and their variances are all 0, and so calculating the correlation of the lag-1 values will involve computing  $0/0 = \text{NaN}$ :

```
[ ]: # plot of lag-1 correlation of PRECT values in col_prect orig dataset
ldcpy.plot(col_prect, "PRECT", sets=["orig", "zfpA1e-7", "zfpA1e-7"], calc="lag1")
```

## Other Spatial Plots

Sometimes, we may want to compute a quantity on the difference between two datasets. For instance, the `zscore` calculation calculates the zscore at each point under the null hypothesis that the true mean is zero, so using the “`calc_of_diff`” `calc_type` calculates the zscore of the diff between two datasets (to find the values that are significantly different between the two datasets). The zscore calculation in particular gives additional information about the percentage of significant gridpoints in the plot title:

```
[ ]: # plot of z-score under null hypothesis that "orig" value= "zfpA1.0" value
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0", "zfpA1e-1"],
    calc="zscore",
    calc_type="calc_of_diff",
    vert_plot=True,
)
```

## Time-Series Plots

We may want to aggregate the data spatially and look for trends over time. Therefore, we can also create a time-series plot of the calculations by changing the `plot_type` to “`time_series`”. For time-series plots, the quantity values are on the y-axis and the x-axis represents time. We are also able to group the data by time, as shown below.

### Basic Time-Series Plot

In the example below, we look at the ‘orig’ TS data in collection `col_ts`, and display the spatial mean at each day of the year (our data consists of 100 days).

```
[ ]: # Time-series plot of TS mean in ds orig dataset
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig"],
    calc="mean",
    plot_type="time_series",
    vert_plot=True,
    legend_loc="best",
)
```

### Using the `group_by` keyword

To group the data by time, use the “`group_by`” keyword. This plot shows the mean standard deviation over all latitude and longitude points for each month. Note that this plot is not too interesting for our sample data, which has only 100 days of data.

```
[ ]: # Time-series plot of TS standard deviation in col_ds "orig" dataset, grouped by month
ldcpy.plot(
    col_ts,
    "TS",
```

(continues on next page)

(continued from previous page)

```

sets=["orig"],
calc="std",
plot_type="time_series",
group_by="time.month",
vert_plot=True,
)

```

One could also group by days of the year, as below. Again, because we have less than a year of data, this plot looks the same as the previous version. However, this approach would be useful with data containing multiple years.

```

[ ]: # Time-series plot of TS mean in col_ts "orig" dataset, grouped by day of year
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0"],
    calc="mean",
    calc_type="calc_of_diff",
    plot_type="time_series",
    group_by="time.dayofyear",
)

```

We can also overlay multiple sets of time-series data on the same plot. For instance, we can plot the mean of two datasets over time. Note that the blue and orange lines overlap almost perfectly:

```

[ ]: # Time-series plot of TS mean in col_ts 'orig' dataset
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0"],
    calc="mean",
    plot_type="time_series",
)

```

If we change the `calc_type` to “diff”, “ratio” or “calc\_of\_diff”, the first element in `sets` is compared against subsequent elements in the `sets` list. For example, we can compare the difference in the mean of two compressed datasets to the original dataset like so:

```

[ ]: # Time-series plot of TS mean differences (with 'orig') in col_ts orig dataset
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig", "zfpA1.0", "zfpA1e-1"],
    calc="mean",
    plot_type="time_series",
    calc_type="diff",
)

```

## Histograms

We can also create a histogram of the data by changing the `plot_type` to ‘`histogram`’. Note that these histograms are currently generated from time-series quantity values (a histogram of spatial values is not currently available).

The histogram below shows the mean values of TS in the ‘`orig`’ and ‘`zfpA1.0`’ datasets in our collection `col_ts`. Recall that this dataset contains 100 timeslices.

```
[ ]: # Histogram of mean TS values in the col_ts (for'orig' and 'zfpA1.0') dataset

ldcpy.plot(col_ts, "TS", sets=["orig", "zfpA1.0"], calc="mean", plot_type="histogram")
```

## Other Time-Series Plots

We can create a periodogram based on a dataset as well, by specifying a `plot_type` of “`periodogram`”.

```
[ ]: ldcpy.plot(col_ts, "TS", sets=["orig"], calc="mean", plot_type="periodogram")

[ ]: ldcpy.plot(col_ts, "TS", sets=["orig"], calc="mean", plot_type="periodogram")
```

## Subsetting

Subsetting is also possible on time-series data. The following plot makes use of the `subset` argument, which is used to plot derived quantities on only a portion of the data. A full list of available subsets is available [here](#). The following plot uses the ‘`first5`’ subset, which only shows the quantity values for the first five time slices (in this case, days) of data:

```
[ ]: # Time-series plot of first five TS standard deviations in col_ts "orig" dataset
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig"],
    calc="std",
    plot_type="time_series",
    subset="first5",
)
```

Additionally, we can specify “`lat`” and “`lon`” keywords for time-series plots that give us a subset of the data at a single point, rather than averaging over all latitudes and longitudes. The nearest latitude and longitude point to the one specified is plotted (and the actual coordinates of the point can be found in the plot title). This plot, for example, shows the difference in mean rainfall between the compressed and original data at the location (44.76, -123.75):

```
[ ]: # Time series plot of PRECT mean data for col_prect "zfpA1e-7" dataset at the location_
↳ (44.76, -123.75)

ldcpy.plot(
    col_prect,
    "PRECT",
    sets=["zfpA1e-7"],
    calc="mean",
    plot_type="time_series",
    lat=44.76,
```

(continues on next page)



(continued from previous page)

```
lon=-123.75,
)
```

```
[ ]: # Time series plot of PRECT mean data for col_prect "zfpA1e-7" dataset at the location_
      ↪ (44.76, -123.75)
```

```
ldcpy.plot(col_prect, "PRECT", sets=["zfpA1e-7"], calc="mean", plot_type="time_series")
```

It is also possible to plot quantities for a subset of the data, by specifying the start and end indices of the data we are interested in. This command creates a time-series plot of the mean TS values for the first 45 days of data:

```
[ ]: # Time series plot of first 45 TS mean data points for col_ts "orig" dataset
```

```
ldcpy.plot(
    col_ts,
    "TS",
    sets=["orig"],
    calc="mean",
    start=0,
    end=44,
    plot_type="time_series",
)
```

```
[ ]:
```

## 2.4.2 CURRENTLY NOT WORKING...needs updating

### 2.4.3 Using data from AWS

A significant amount of Earth System Model (ESM) data is publicly available online, including data from the CESM Large Ensemble, CMIP5, and CMIP6 datasets. For accessing a single file, we can specify the file (typically netcdf or zarr format) and its location and then use fsspec (the “Filesystem Spec+ python package) and xarray to create a `array.dataset`. For several files, the `intake_esm` python module (<https://github.com/intake/intake-esm>) is particularly nice for obtaining the data and put it into an `xarray.dataset`.

This notebook assumes familiarity with the Tutorial Notebook. It additionally shows how to gather data from an ESM collection, put it into a dataset, and then create simple plots using the data with `ldcpy`.

#### Example Data

The example data we use is from the CESM Large Ensemble, member 31. This ensemble data has been lossily compressed and reconstructed as part of a blind evaluation study of lossy data compression in LENS (e.g., <http://www.cesm.ucar.edu/projects/community-projects/LENS/projects/lossy-data-compression.html> or <https://gmd.copernicus.org/articles/9/4381/2016/>).

Most of the data from the CESM Large Ensemble Project has been made available on Amazon Web Services (Amazon S3), see [http://ncar-aws-www.s3-website-us-west-2.amazonaws.com/CESM\\_LENS\\_on\\_AWS.htm](http://ncar-aws-www.s3-website-us-west-2.amazonaws.com/CESM_LENS_on_AWS.htm).

For comparison purposes, the original (non-compressed) data for Ensemble 31 has recently been made available on Amazon Web Services (Amazon S3) in the “ncar-cesm-lens-baker-lossy-compression-test” bucket.

```
[1]: # Add ldcpy root to system path
import sys

sys.path.insert(0, '../..../')

# Import ldcpy package
# Autoreloads package everytime the package is called, so changes to code will be
# reflected in the notebook if the above sys.path.insert(...) line is uncommented.
%load_ext autoreload
%autoreload 2
import fsspec
import intake
import xarray as xr

import ldcpy

# display the plots in this notebook
%matplotlib inline

# silence warnings
import warnings

warnings.filterwarnings("ignore")
```

### Method 1: using fsspec and xr.open\_zarr

First, specify the filesystem and location of the data. Here we are accessing the original data from CESM-LENS ensemble 31, which is available on Amazon S3 in the store named “ncar-cesm-lens-baker-lossy-compression-test” bucket.

First we listing all available files (which are timeseries files containing a single variable) for that dataset. Note that unlike in the TutorialNotebook (which used NetCDF files), these files are all zarr format. Both monthly and daily data is available.

```
[3]: fs = fsspec.filesystem("s3", anon=True)
stores = fs.ls("ncar-cesm-lens-baker-lossy-compression-test/lens-ens31/")[1:]
stores[:]

-----
CancelledError                                Traceback (most recent call last)
/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/connector.py
  → in _wrap_create_connection(self, req, timeout, client_error, *args, **kwargs)
      968         with CeilTimeout(timeout.sock_connect):
--> 969         return await self._loop.create_connection(*args, **kwargs) #
  → type: ignore # noqa
      970         except cert_errors as exc:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/asyncio/base_events.py in create_
  → connection(self, protocol_factory, host, port, ssl, family, proto, flags, sock, local_
  → addr, server_hostname, ssl_handshake_timeout)
      948         logger.debug("connect %r to %r", sock, address)
--> 949         await self.sock_connect(sock, address)
```

(continues on next page)

(continued from previous page)

```

950             except OSError as exc:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/asyncio/selector_events.py in sock_
↳ connect(self, sock, address)
    472         self._sock_connect(fut, sock, address)
--> 473         return await fut
    474

CancelledError:

During handling of the above exception, another exception occurred:

TimeoutError                                Traceback (most recent call last)
/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/client.py in _
↳ request(self, method, str_or_url, params, data, json, cookies, headers, skip_auto_
↳ headers, auth, allow_redirects, max_redirects, compress, chunked, expect100, raise_for_
↳ status, read_until_eof, proxy, proxy_auth, timeout, verify_ssl, fingerprint, ssl_
↳ context, ssl, proxy_headers, trace_request_ctx, read_bufsize)
    520             conn = await self._connector.connect(
--> 521                 req, traces=traces, timeout=real_timeout
    522             )

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/connector.py_
↳ in connect(self, req, traces, timeout)
    534         try:
--> 535             proto = await self._create_connection(req, traces, timeout)
    536             if self._closed:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/connector.py_
↳ in _create_connection(self, req, traces, timeout)
    891         else:
--> 892             _, proto = await self._create_direct_connection(req, traces, timeout)
    893

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/connector.py_
↳ in _create_direct_connection(self, req, traces, timeout, client_error)
    1031             req=req,
-> 1032             client_error=client_error,
    1033         )

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/connector.py_
↳ in _wrap_create_connection(self, req, timeout, client_error, *args, **kwargs)
    968         with CeilTimeout(timeout.sock_connect):
--> 969             return await self._loop.create_connection(*args, **kwargs) #_
↳ type: ignore # noqa
    970         except cert_errors as exc:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/async_timeout/__init__
↳ .py in __exit__(self, exc_type, exc_val, exc_tb)
    44         exc_tb: TracebackType) -> Optional[bool]:
--> 45         self._do_exit(exc_type)
    46         return None

```

(continues on next page)

(continued from previous page)

```

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/async_timeout/__init__
↳.py in _do_exit(self, exc_type)
    91             self._task = None
--> 92             raise asyncio.TimeoutError
    93             if self._timeout is not None and self._cancel_handler is not None:

TimeoutError:

The above exception was the direct cause of the following exception:

ServerTimeoutError                                Traceback (most recent call last)
/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/fsspec/asyn.py in _
↳runner(event, coro, result, timeout)
    24         try:
--> 25             result[0] = await coro
    26         except Exception as ex:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/s3fs/core.py in _
↳ls(self, path, detail, refresh)
    794         else:
--> 795             files = await self._lsdir(path, refresh)
    796             if not files and "/" in path:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/s3fs/core.py in _
↳lsdir(self, path, refresh, max_items, delimiter, prefix)
    577             dircache = []
--> 578             async for i in it:
    579                 dircache.extend(i.get("CommonPrefixes", []))

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/paginate.
↳py in __anext__(self)
    31         while True:
--> 32             response = await self._make_request(current_kwargs)
    33             parsed = self._extract_parsed_response(response)

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/client.py
↳in _make_api_call(self, operation_name, api_params)
    141             http, parsed_response = await self._make_request(
--> 142                 operation_model, request_dict, request_context)
    143

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/client.py
↳in _make_request(self, operation_model, request_dict, request_context)
    160         try:
--> 161             return await self._endpoint.make_request(operation_model, request_
↳dict)
    162         except Exception as e:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/endpoint.
↳py in _send_request(self, request_dict, operation_model)
    82             request_dict, success_response,

```

(continues on next page)

(continued from previous page)

```

---> 83                                     exception):
      84             attempts += 1

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/endpoint.
↳py in _needs_retry(self, attempts, operation_model, request_dict, response, caught_
↳exception)
      216             operation=operation_model, attempts=attempts,
--> 217             caught_exception=caught_exception, request_dict=request_dict)
      218             handler_response = first_non_none_response(responses)

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/hooks.py_
↳in _emit(self, event_name, kwargs, stop_on_response)
      28             else:
---> 29                 response = handler(**kwargs)
      30

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in __call__(self, attempts, response, caught_exception, **kwargs)
      182             """
--> 183             if self._checker(attempts, response, caught_exception):
      184                 result = self._action(attempts=attempts)

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in __call__(self, attempt_number, response, caught_exception)
      250             should_retry = self._should_retry(attempt_number, response,
--> 251                                     caught_exception)
      252             if should_retry:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in _should_retry(self, attempt_number, response, caught_exception)
      276             # propogate if one has occurred.
--> 277             return self._checker(attempt_number, response, caught_exception)
      278

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in __call__(self, attempt_number, response, caught_exception)
      316             checker_response = checker(attempt_number, response,
--> 317                                     caught_exception)
      318             if checker_response:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in __call__(self, attempt_number, response, caught_exception)
      222             return self._check_caught_exception(
--> 223                 attempt_number, caught_exception)
      224             else:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/botocore/retryhandler.
↳py in _check_caught_exception(self, attempt_number, caught_exception)
      358             # then this exception just propogates out past the retry code.
--> 359             raise caught_exception

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiobotocore/endpoint.
↳py in _do_get_response(self, request, operation_model)

```

(continues on next page)

(continued from previous page)

```

147         if http_response is None:
--> 148             http_response = await self._send(request)
149     except aiohttp.ClientConnectionError as e:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/core.py in _send(self, request)
->py in _send(self, request)
229     async def _send(self, request):
--> 230         return await self.http_session.send(request)
231

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/core.py
->http_session.py in send(self, request)
155         resp = await self._session.request(
--> 156             request.method, url=url, headers=headers_, data=data,
->proxy=proxy_url)
157

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/aiohttp/client.py in _
->request(self, method, str_or_url, params, data, json, cookies, headers, skip_auto_
->headers, auth, allow_redirects, max_redirects, compress, chunked, expect100, raise_for_
->status, read_until_eof, proxy, proxy_auth, timeout, verify_ssl, fingerprint, ssl_
->context, ssl, proxy_headers, trace_request_ctx, read_bufsize)
525             "Connection timeout " "to host {}".format(url)
--> 526         ) from exc
527

ServerTimeoutError: Connection timeout to host https://ncar-cesm-lens-baker-lossy-
->compression-test.s3.amazonaws.com/?list-type=2&prefix=lens-ens31%2F&delimiter=%2F&
->encoding-type=url

```

The above exception was the direct cause of the following exception:

```

FSTimeoutError                                Traceback (most recent call last)
/glade/scratch/abaker/ipykernel_13953/1112949793.py in <module>
      1 fs = fsspec.filesystem("s3", anon=True)
----> 2 stores = fs.ls("ncar-cesm-lens-baker-lossy-compression-test/lens-ens31/")[1:]
      3 stores[:]

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/fsspec/asyn.py in _
->wrapper(*args, **kwargs)
    89     def wrapper(*args, **kwargs):
    90         self = obj or args[0]
--> 91         return sync(self.loop, func, *args, **kwargs)
    92
    93     return wrapper

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/fsspec/asyn.py in _
->sync(loop, func, timeout, *args, **kwargs)
    67     if isinstance(return_result, asyncio.TimeoutError):
    68         # suppress asyncio.TimeoutError, raise FSTimeoutError
--> 69         raise FSTimeoutError from return_result
    70     elif isinstance(return_result, BaseException):

```

(continues on next page)

(continued from previous page)

```
71         raise return_result
```

```
FSTimeoutError:
```

Then we select the file from the store that we want and open it as an `xarray.Dataset` using `xr.open_zarr()`. Here we grab data for the first 2D daily variable, FLNS (net longwave flux at surface, in  $W/m^2$ ), in the list (accessed by its location – `stores[0]`).

```
[ ]: store = fs.get_mapper(stores[0])
     ds_flns = xr.open_zarr(store, consolidated=True)
     ds_flns
```

The above returned an `xarray.Dataset`.

Now let's grab the TMQ (Total vertically integrated precipitable water) and the TS (surface temperature data) and PRECT (precipitation rate) data from AWS.

```
[ ]: # TMQ data
     store2 = fs.get_mapper(stores[16])
     ds_tmq = xr.open_zarr(store2, consolidated=True)
     # TS data
     store3 = fs.get_mapper(stores[20])
     ds_ts = xr.open_zarr(store3, consolidated=True)
     # PRECT data
     store4 = fs.get_mapper(stores[11])
     ds_prect = xr.open_zarr(store4, consolidated=True)
```

Now we have the original data for FLNS and TMQ and TS and PRECT. Next we want to get the lossy compressed variants to compare with these.

## Method 2: Using intake\_esm

Now we will demonstrate using the `intake_esm` module to get the lossy variants of the files retrieved above. We can use the `intake_esm` module to search for and open several files as `xarray.Dataset` objects. The code below is modified from the `intake_esm` documentation, available here: <https://intake-esm.readthedocs.io/en/latest/?badge=latest#overview>.

We want to use ensemble 31 data from the CESM-LENS collection on AWS, which (as explained above) has been subjected to lossy compression. Many catalogs for publicly available datasets are accessible via `intake-esm` can be found at <https://github.com/NCAR/intake-esm-datastore/tree/master/catalogs>, including for CESM-LENS. We can open that collection as follows (see here: <https://github.com/NCAR/esm-collection-spec/blob/master/collection-spec/collection-spec.md#attribute-object>):

```
[2]: aws_loc = (
     "https://raw.githubusercontent.com/NCAR/cesm-lens-aws/master/intake-catalogs/aws-
     ↪cesm1-le.json"
     )
     aws_col = intake.open_esm_datastore(aws_loc)
     aws_col

-----
KeyboardInterrupt                                Traceback (most recent call last)
/glade/scratch/abaker/ipykernel_13953/875294884.py in <module>
      2     "https://raw.githubusercontent.com/NCAR/cesm-lens-aws/master/intake-catalogs/
     ↪aws-cesm1-le.json"
```

(continues on next page)

(continued from previous page)

```

3 )
----> 4 aws_col = intake.open_esm_datastore(aws_loc)
5 aws_col

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/intake_esm/core.py in
->__init__(self, esmcol_obj, esmcol_data, progressbar, sep, csv_kwargs, **kwargs)
81     super(esm_datastore, self).__init__(**kwargs)
82     if isinstance(esmcol_obj, (str, pathlib.PurePath)):
--> 83         self.esmcol_data, self.esmcol_path = _fetch_and_parse_json(esmcol_
->obj)
84         self._df, self.catalog_file = _fetch_catalog(self.esmcol_data,
->esmcol_obj, csv_kwargs)
85

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/intake_esm/utils.py
->in _fetch_and_parse_json(input_path)
48
49     try:
--> 50         if _is_valid_url(input_path):
51             resp = requests.get(input_path)
52             data = resp.json()

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/intake_esm/utils.py
->in _is_valid_url(url)
27         and result.netloc
28         and result.path
--> 29         and (requests.get(url).status_code == 200)
30     )
31     except Exception:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/requests/api.py in
->get(url, params, **kwargs)
73     """
74
--> 75     return request('get', url, params=params, **kwargs)
76
77

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/requests/api.py in
->request(method, url, **kwargs)
59     # cases, and look like a memory leak in others.
60     with sessions.Session() as session:
--> 61         return session.request(method=method, url=url, **kwargs)
62
63

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/requests/sessions.py
->in request(self, method, url, params, data, headers, cookies, files, auth, timeout,
->allow_redirects, proxies, hooks, stream, verify, cert, json)
540     }
541     send_kwargs.update(settings)
--> 542     resp = self.send(prepare, **send_kwargs)

```

(continues on next page)



(continued from previous page)

```

543
544         return resp

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/requests/sessions.py
↳in send(self, request, **kwargs)
    653
    654         # Send the request
--> 655         r = adapter.send(request, **kwargs)
    656
    657         # Total elapsed time of the request (approximately)

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/requests/adapters.py
↳in send(self, request, stream, timeout, verify, cert, proxies)
    447             decode_content=False,
    448             retries=self.max_retries,
--> 449             timeout=timeout
    450         )
    451

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/
↳connectionpool.py in urlopen(self, method, url, body, headers, retries, redirect,
↳assert_same_host, timeout, pool_timeout, release_conn, chunked, body_pos, **response_
↳kw)
    704             body=body,
    705             headers=headers,
--> 706             chunked=chunked,
    707         )
    708

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/
↳connectionpool.py in _make_request(self, conn, method, url, timeout, chunked,
↳**httplib_request_kw)
    380         # Trigger any extra validation we need to do.
    381         try:
--> 382             self._validate_conn(conn)
    383         except (SocketTimeout, BaseSSLError) as e:
    384             # Py2 raises this as a BaseSSLError, Py3 raises it as socket timeout.

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/
↳connectionpool.py in _validate_conn(self, conn)
    1008         # Force connect early to allow us to validate the connection.
    1009         if not getattr(conn, "sock", None): # AppEngine might not have `.sock`
-> 1010             conn.connect()
    1011
    1012         if not conn.is_verified:

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/connection.py
↳in connect(self)
    356     def connect(self):
    357         # Add certificate verification
--> 358         conn = self._new_conn()
    359         hostname = self.host

```

(continues on next page)

(continued from previous page)

```

360         tls_in_tls = False

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/connection.py
-> in _new_conn(self)
    173         try:
    174             conn = connection.create_connection(
--> 175                 (self._dns_host, self.port), self.timeout, **extra_kw
    176             )
    177

/ncar/usr/jupyterhub/envs/cmip6-201910/lib/python3.7/site-packages/urllib3/util/
-> connection.py in create_connection(address, timeout, source_address, socket_options)
    84         if source_address:
    85             sock.bind(source_address)
--> 86         sock.connect(sa)
    87         return sock
    88

KeyboardInterrupt:

```

Next, we search for the subset of the collection (dataset and variables) that we are interested in. Let's grab FLNS, TMQ, and TS daily data from the atm component for our comparison (available data in this collection is listed here: [http://ncar-aws-www.s3-website-us-west-2.amazonaws.com/CESM\\_LENS\\_on\\_AWS.htm](http://ncar-aws-www.s3-website-us-west-2.amazonaws.com/CESM_LENS_on_AWS.htm)).

```

[ ]: # we want daily data for FLNS, TMQ, and TS and PRECT
aws_col_subset = aws_col.search(
    component="atm",
    frequency="daily",
    experiment="20C",
    variable=["FLNS", "TS", "TMQ", "PRECT"],
)
# display header info to verify that we got the right variables
aws_col_subset.df.head()

```

Then we load matching catalog entries into xarray datasets ([https://intake-esm.readthedocs.io/en/latest/api.html#intake\\_esm.core.esm\\_datastore.to\\_dataset\\_dict](https://intake-esm.readthedocs.io/en/latest/api.html#intake_esm.core.esm_datastore.to_dataset_dict)). We create a dictionary of datasets:

```

[ ]: dset_dict = aws_col_subset.to_dataset_dict(
    zarr_kwargs={"consolidated": True, "decode_times": True},
    storage_options={"anon": True},
    cdf_kwargs={"chunks": {}, "decode_times": False},
)
dset_dict

```

Check the dataset keys to ensure that what we want is present. Here we only have one entry in the dictionary as we requested the same time period and output frequency for all variables:

```

[ ]: dset_dict.keys()

```

Finally, put the dataset that we are interested from the dictionary into its own dataset variable. (We want the 20th century daily data – which is our only option.)

Also note from above that there are 40 ensemble members - and we just want ensemble 31 (member\_id = 30 as can be

seen in the coordinates above).

```
[ ]: aws_ds = dset_dict["atm.20C.daily"]
aws_ds = aws_ds.isel(member_id=30)
aws_ds
```

Now we have datasets for the original and the lossy compressed data for FLNS, TMQ, PRECT, and TS, which we can extract into a dataset for each variable:

```
[ ]: # extract the three variables from aws_ds as datasets
aws_flns = aws_ds["FLNS"].to_dataset()
aws_tmq = aws_ds["TMQ"].to_dataset()
aws_ts = aws_ds["TS"].to_dataset()
aws_prect = aws_ds["PRECT"].to_dataset()
```

## 2.4.4 Use *ldcpy* to compare the original data to the lossy compressed data

To use *ldcpy*, we need to group the data that we want to compare (like variables) into dataset collections. In the Tutorial Notebook, we used *ldcpy.open\_datasets()* to do this as we needed to get the data from the NetCDF files. Here we already loaded the data from AWS into datasets, so we just need to use *ldcpy.collect\_datasets()* to form collections of the datasets that we want to compare.

*ldcpy.collect\_datasets()* requires the following three arguments:

- *varnames* : the variable(s) of interest to combine across files (typically the timeseries file variable name)
- *list\_of\_ds* : a list of the xarray datasets
- *labels* : a corresponding list of names (or labels) for each dataset in the collection

Note: This function is a wrapper for *xarray.concat()*, and any additional key/value pairs passed in as a dictionary are used as arguments to *xarray.concat()*.

We will create 4 collections for *ldcpy* (one each for FLNS, TMQ, PRECT, and TS) and assign labels “original” and “lossy” to the respective datasets.

```
[ ]: # FLNS collection
col_flns = ldcpy.collect_datasets("cam-fv", ["FLNS"], [ds_flns, aws_flns], ["original",
↪ "lossy"])
col_flns
```

```
[ ]: # TMQ collection
col_tmq = ldcpy.collect_datasets("cam-fv", ["TMQ"], [ds_tmq, aws_tmq], ["original",
↪ "lossy"])
```

```
[ ]: # Ts collection
col_ts = ldcpy.collect_datasets("cam-fv", ["TS"], [ds_ts, aws_ts], ["original", "lossy"])
```

```
[ ]: # PRECT collection
col_prect = ldcpy.collect_datasets(
    "cam-fv", ["PRECT"], [ds_prect, aws_prect], ["original", "lossy"]
)
```

Now that we have our collections, we can do some comparisons. Note that these are large files, so make sure you have sufficient compute/memory.

```
[ ]: # Time-series plot of PRECT mean in col_ds 'original' dataset - first 100 daysa
ldcpy.plot(
    col_prect,
    "PRECT",
    sets=["original", "lossy"],
    calc="mean",
    plot_type="time_series",
    start=0,
    end=100,
)
```

```
[ ]: # print statistics about 'original', 'lossy', and diff between the two datasets for TMQ at_
↳time slice 365
ldcpy.compare_stats(col_tmq.isel(time=365), "TMQ", ["original", "lossy"])
```

The original data for FLNS and TMQ and TS and PRECT (from Amazon S3 in the “ncar-cesm-lens-baker-lossy-compression-test” bucket) was loaded above using method 1. An alternative would be to create our own catalog for this data for use with intake-esm. To illustrate this, we created a `test_catalog.csv` and `test_collection.json` file for this particular simple example.

We first open our collection.

```
[ ]: my_col_loc = "./collections/test_collection.json"
my_col = intake.open_esm_datastore(my_col_loc)
my_col
```

```
[ ]: # printing the head() gives us the file names
my_col.df.head()
```

Let’s load all of these into our dictionary! (So we don’t need to do the search to make a subset of variables as above in Method 2.)

```
[ ]: my_dset_dict = my_col.to_dataset_dict(
    zarr_kwargs={"consolidated": True, "decode_times": True},
    storage_options={"anon": True},
)
my_dset_dict
```

```
[ ]: # we again just want the 20th century daily data
my_ds = my_dset_dict["atm.20C.daily"]
my_ds
```

Now we can make a dataset for each variable as before.

```
[ ]: my_ts = my_ds["TS"].to_dataset()
my_tmq = my_ds["TMQ"].to_dataset()
my_prect = my_ds["PRECT"].to_dataset()
my_flns = my_ds["FLNS"].to_dataset()
```

And now we can form new collections as before and do comparisons...

```
[ ]:
```

## 2.4.5 NCAR JupyterHub Large Data Example Notebook

*Note: If you do not have access to the NCAR machine, please look at the AWS-LENS example notebook instead.*

This notebook demonstrates how to compare large datasets on glade with ldcpy. In particular, we will look at data from CESM-LENS1 project (<http://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html>). In doing so, we will start a DASK client from Jupyter. This notebook is meant to be run on NCAR's JupyterHub (<https://jupyterhub.hpc.ucar.edu>). We will use a subset of the CESM-LENS1 data on glade is located in `/glade/p/cisl/asap/ldcpy_sample_data/lens`.

We assume that you have a copy of the ldcpy code on NCAR's glade filesystem, obtained via: `git clone https://github.com/NCAR/ldcpy.git`

When you launch your NCAR JupyterHub session, you will need to indicate a machine and then you will need your charge account. You can then launch the session and navigate to this notebook.

NCAR's JupyterHub documentation: <https://www2.cisl.ucar.edu/resources/jupyterhub-ncar>

Here's another good resource for using NCAR's JupyterHub: <https://ncar-hackathons.github.io/jupyterlab-tutorial/jhub.html>

**You need to run your notebook with the “cmip6-2019.10” kernel (choose from the dropdown in the upper left.)**

Note that the compressed data that we are using was generated for this paper:

Allison H. Baker, Dorit M. Hammerling, Sheri A. Mickelson, Haiying Xu, Martin B. Stolpe, Phillipe Naveau, Ben Sanderson, Imme Ebert-Uphoff, Savini Samarasinghe, Francesco De Simone, Francesco Carbone, Christian N. Gencarelli, John M. Dennis, Jennifer E. Kay, and Peter Lindstrom, “Evaluating Lossy Data Compression on Climate Simulation Data within a Large Ensemble.” *Geoscientific Model Development*, 9, pp. 4381-4403, 2016 (<https://gmd.copernicus.org/articles/9/4381/2016/>)

### Setup

Let's set up our environment. First, make sure that you are using the cmip6-2019.10 kernel. Then you will need to modify the path below to indicate where you have cloned ldcpy. (*Note: soon we will install ldcpy on Cheyenne/Casper in the cmip6-2019.10 kernel .*)

If you want to use the dask dashboard, then the dask.config link must be set below (modify for your path in your browser).

```
[1]: # Make sure you are using the cmip6-2019.10 kernel on NCAR JupyterHub

# Add ldcpy root to system path (MODIFY FOR YOUR LDCPY CODE LOCATION)
import sys

sys.path.insert(0, '/glade/u/home/abaker/repos/ldcpy')
import ldcpy

# Display output of plots directly in Notebook
%matplotlib inline

# Automatically reload module if it is edited
%reload_ext autoreload
%autoreload 2

# silence warnings
import warnings
```

(continues on next page)

(continued from previous page)

```
warnings.filterwarnings("ignore")

# if you want to use the DASK dashboard, then modify the below and run
# import dask
# dask.config.set(
#     {'distributed.dashboard.link': 'https://jupyterhub.hpc.ucar.edu/ch/user/abaker/proxy/
↪ {port}/status'}
# )
```

### Connect to DASK distributed cluster :

For NCAR machines, we can use `ncar_jobqueue` package (instead of the `dask_jobqueue` package). (The cluster object is for a single compute node: <https://jobqueue.dask.org/en/latest/howitworks.html>)

```
[2]: from dask.distributed import Client
from ncar_jobqueue import NCARCluster

cluster = NCARCluster(project='NTDD0004')
# scale as needed
cluster.adapt(minimum_jobs=1, maximum_jobs=30)
cluster
client = Client(cluster)
```

The scheduler creates a normal-looking job script that it can submit multiple times to the queue:

```
[3]: # Look at the job script (optional)
print(cluster.job_script())

#!/usr/bin/env bash

#PBS -N dask-worker-casper-dav
#PBS -q casper
#PBS -A NTDD0004
#PBS -l select=1:ncpus=1:mem=25GB
#PBS -l walltime=01:00:00
#PBS -e /glade/scratch/abaker/dask/casper-dav/logs/
#PBS -o /glade/scratch/abaker/dask/casper-dav/logs/

/ncar/usr/jupyterhub/envs/cmip6-201910/bin/python -m distributed.cli.dask_worker tcp://
↪ 10.12.206.63:39183 --nthreads 2 --memory-limit 23.28GiB --name dummy-name --nanny --
↪ death-timeout 60 --local-directory /glade/scratch/abaker/dask/casper-dav/local-dir --
↪ interface ib0 --protocol tcp://
```

## The sample data on the glade filesystem

In /glade/p/cisl/asap/ldcpy\_sample\_data/lens on glade, we have TS (surface temperature), PRECT (precipitation rate), and PS (surface pressure) data from CESM-LENS1. These all are 2D variables. TS and PRECT have daily output, and PS has monthly output. We have the compressed and original versions of all these variables that we would like to compare with ldcpy.

First we list what is in this directory (two subdirectories):

```
[4]: # list directory contents
import os

os.listdir("/glade/p/cisl/asap/ldcpy_sample_data/lens")
```

```
[4]: ['README.txt', 'lossy', 'orig']
```

Now we look at the contents of each subdirectory. We have 14 files in each, consisting of 2 different timeseries files for each variable (1920-2005 and 2006-2080).

```
[5]: # list lossy directory contents (files that have been lossy compressed and reconstructed)
lossy_files = os.listdir("/glade/p/cisl/asap/ldcpy_sample_data/lens/lossy")
lossy_files
```

```
[5]: ['c.FLUT.daily.20060101-20801231.nc',
      'c.CCN3.monthly.192001-200512.nc',
      'c.FLNS.monthly.192001-200512.nc',
      'c.LHFLX.daily.20060101-20801231.nc',
      'c.TS.daily.19200101-20051231.nc',
      'c.SHFLX.daily.20060101-20801231.nc',
      'c.U.monthly.200601-208012.nc',
      'c.TREFHT.monthly.192001-200512.nc',
      'c.LHFLX.daily.19200101-20051231.nc',
      'c.TMQ.monthly.192001-200512.nc',
      'c.PRECT.daily.20060101-20801231.nc',
      'c.PS.monthly.200601-208012.nc',
      'c.Z500.daily.20060101-20801231.nc',
      'c.FLNS.monthly.200601-208012.nc',
      'c.Q.monthly.192001-200512.nc',
      'c.U.monthly.192001-200512.nc',
      'c.CLOUD.monthly.192001-200512.nc',
      'c.so4_a2_SRF.daily.20060101-20801231.nc',
      'c.PRECT.daily.19200101-20051231.nc',
      'c.TAUX.daily.20060101-20801231.nc',
      'c.CLOUD.monthly.200601-208012.nc',
      'c.TS.daily.20060101-20801231.nc',
      'c.FSNTOA.daily.20060101-20801231.nc',
      'c.Q.monthly.200601-208012.nc',
      'c.CCN3.monthly.200601-208012.nc',
      'c.TREFHT.monthly.200601-208012.nc',
      'c.TMQ.monthly.200601-208012.nc',
      'c.PS.monthly.192001-200512.nc']
```

```
[6]: # list orig (i.e., uncompressed) directory contents
orig_files = os.listdir("/glade/p/cisl/asap/ldcpy_sample_data/lens/orig")
orig_files
```

```
[6]: ['CLOUD.monthly.200601-208012.nc',
      'LHFLX.daily.20060101-20801231.nc',
      'so4_a2_SRF.daily.20060101-20801231.nc',
      'PRECT.daily.20060101-20801231.nc',
      'TMQ.monthly.192001-200512.nc',
      'TAUX.daily.20060101-20801231.nc',
      'CCN3.monthly.200601-208012.nc',
      'PS.monthly.192001-200512.nc',
      'FLUT.daily.20060101-20801231.nc',
      'PRECT.daily.19200101-20051231.nc',
      'TS.daily.20060101-20801231.nc',
      'PS.monthly.200601-208012.nc',
      'TS.daily.19200101-20051231.nc',
      'Z500.daily.20060101-20801231.nc',
      'SHFLX.daily.20060101-20801231.nc',
      'Q.monthly.200601-208012.nc',
      'TMQ.monthly.200601-208012.nc',
      'U.monthly.200601-208012.nc',
      'FSNTOA.daily.20060101-20801231.nc',
      'TREFHT.monthly.200601-208012.nc',
      'FLNS.monthly.192001-200512.nc',
      'FLNS.monthly.200601-208012.nc']
```

We can look at how big these files are...

```
[7]: print("Original files")
      for f in orig_files:
          print(
              f,
              " ",
              os.stat("/glade/p/cisl/asap/ldcpy_sample_data/lens/orig/" + f).st_size / 1e9,
              "GB",
          )
```

```
Original files
CLOUD.monthly.200601-208012.nc    3.291160567 GB
LHFLX.daily.20060101-20801231.nc  4.804299285 GB
so4_a2_SRF.daily.20060101-20801231.nc  4.755841518 GB
PRECT.daily.20060101-20801231.nc  4.909326733 GB
TMQ.monthly.192001-200512.nc     0.160075734 GB
TAUX.daily.20060101-20801231.nc  4.856588097 GB
CCN3.monthly.200601-208012.nc    4.053932835 GB
PS.monthly.192001-200512.nc     0.12766304 GB
FLUT.daily.20060101-20801231.nc  4.091392046 GB
PRECT.daily.19200101-20051231.nc  5.629442994 GB
TS.daily.20060101-20801231.nc    3.435295036 GB
PS.monthly.200601-208012.nc     0.111186435 GB
TS.daily.19200101-20051231.nc    3.962086636 GB
Z500.daily.20060101-20801231.nc  3.244290942 GB
SHFLX.daily.20060101-20801231.nc  4.852980899 GB
Q.monthly.200601-208012.nc      3.809397495 GB
TMQ.monthly.200601-208012.nc     0.139301278 GB
U.monthly.200601-208012.nc      4.252600112 GB
```

(continues on next page)



(continued from previous page)

```

FSNTOA.daily.20060101-20801231.nc    4.101201335 GB
TREFHT.monthly.200601-208012.nc      0.110713886 GB
FLNS.monthly.192001-200512.nc        0.170014618 GB
FLNS.monthly.200601-208012.nc        0.148417532 GB

```

## Open datasets

First, let's look at the original and reconstructed files for the monthly surface Pressure (PS) data for 1920-2006. We begin by using `ldcpy.open_dataset()` to open the files of interest into our dataset collection. Usually we want chunks to be 100-150MB, but this is machine and app dependent.

[8]: *# load the first 86 years of montly surface pressure into a collection*

```

col_PS = ldcpy.open_datasets(
    "cam-fv",
    ["PS"],
    [
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/orig/PS.monthly.192001-200512.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/lossy/c.PS.monthly.192001-200512.nc",
    ],
    ["orig", "lossy"],
    chunks={"time": 500},
)
col_PS

```

dataset size in GB 0.46

```

[8]: <xarray.Dataset>
Dimensions:      (collection: 2, time: 1032, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 1920-02-01 00:00:00 ... 2006-01-01 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288),
↳ meta=np.ndarray>
  * collection   (collection) <U5 'orig' 'lossy'
Data variables:
  PS            (collection, time, lat, lon) float32 dask.array<chunksize=(1, 500, 192,
↳ 288), meta=np.ndarray>
Attributes: (12/14)
  Conventions:    CF-1.0
  source:         CAM
  case:           b.e11.B20TRC5CNBDRD.f09_g16.031
  title:          UNSET
  logname:        mickelso
  host:           ys0219
  ...
  initial_file:   b.e11.B20TRC5CNBDRD.f09_g16.001.cam.i.1920-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:        Tue Nov  3 13:51:10 2020: ncks -L 5 PS.monthly.192001-2...
  NCO:            netCDF Operators version 4.7.9 (Homepage = http://nco.s...

```

(continues on next page)

(continued from previous page)

```
cell_measures:    area: cell_area
data_type:        cam-fv
```

## Data comparison

Now we use the ldcpy package features to compare the data.

## Surface Pressure

Let's look at the comparison statistics at the first timeslice for PS.

```
[21]: ps0 = col_PS.isel(time=0)
ldcpy.compare_stats(ps0, "PS", ["orig", "lossy"])
```

```
<IPython.core.display.HTML object>
```

	orig	lossy
mean	98509	98493
variance	8.4256e+07	8.425e+07
standard deviation	9179.1	9178.8
min value	51967	51952
max value	1.0299e+05	1.0298e+05
probability positive	1	1
number of zeros	0	0
spatial autocorr - latitude	0.98434	0.98434
spatial autocorr - longitude	0.99136	0.99136
entropy estimate	0.40644	0.11999

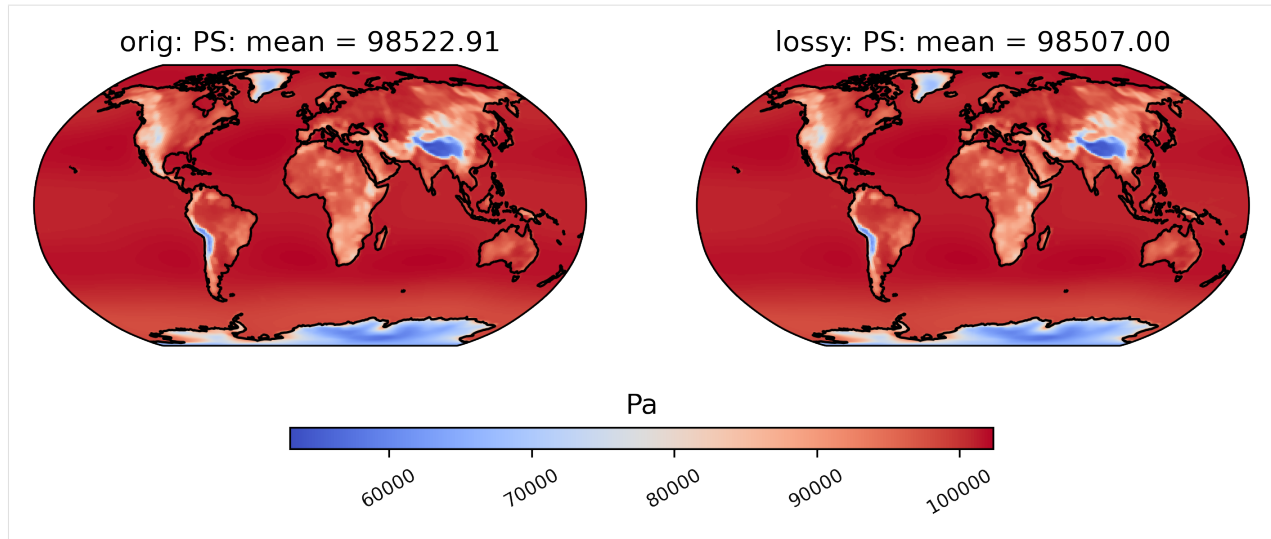
```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

	lossy
max abs diff	31.992
min abs diff	0
mean abs diff	15.906
mean squared diff	253.01
root mean squared diff	18.388
normalized root mean squared diff	0.0003587
normalized max pointwise error	0.00062698
pearson correlation coefficient	1
ks p-value	1.6583e-05
spatial relative error(% > 0.0001)	69.085
max spatial relative error	0.00048247
Data SSIM	0.91512

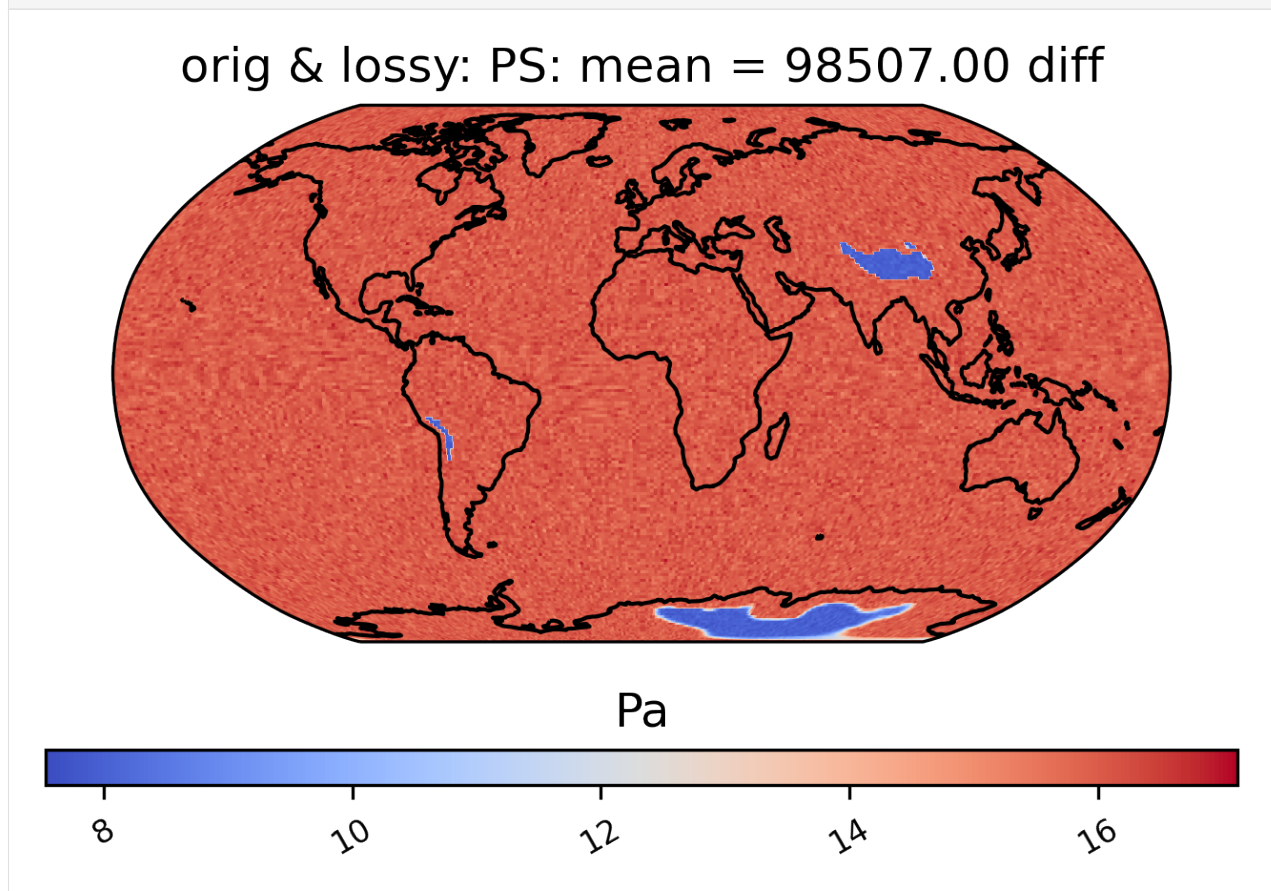
Now we make a plot to compare the mean PS values across time in the orig and lossy datasets.

```
[22]: # comparison between mean PS values (over the 86 years) in col_PS orig and lossy datasets
ldcpy.plot(col_PS, "PS", sets=["orig", "lossy"], calc="mean")
```



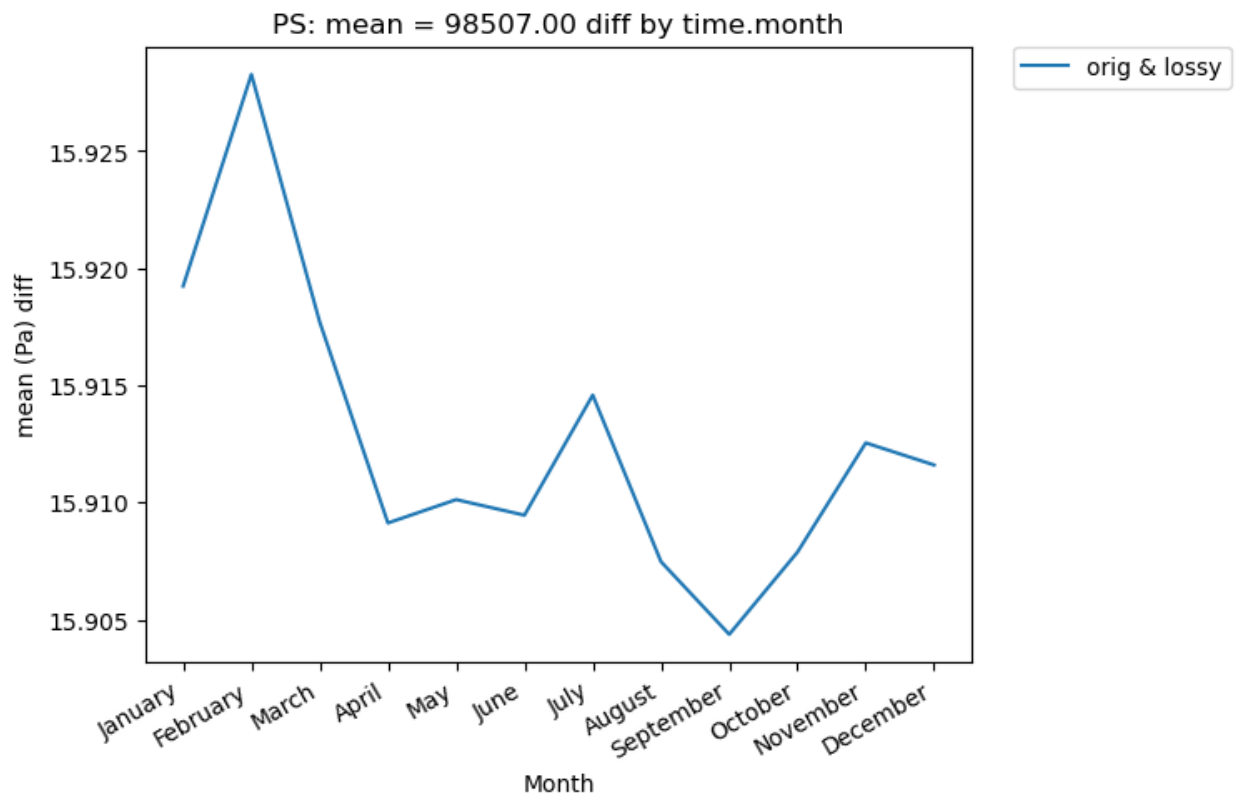
Now we instead show the difference plot for the above plots.

```
[23]: # diff between mean PS values in col_PS orig and lossy datasets
ldcpy.plot(col_PS, "PS", sets=["orig", "lossy"], calc="mean", calc_type="diff")
```

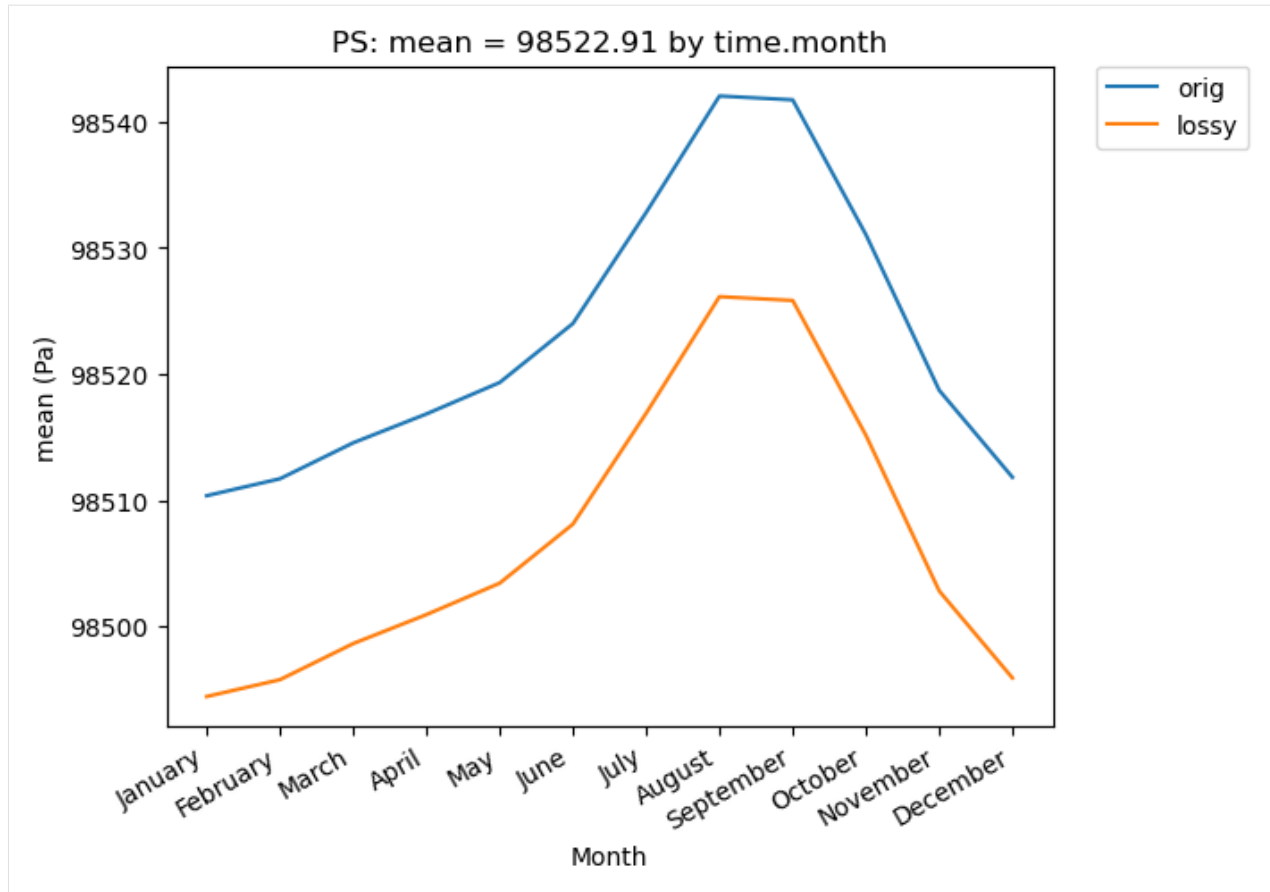


We can also look at mean differences over time. Here we are looking at the spatial averages and then grouping by day of the year. If doing a timeseries plot for this much data, using “group\_by” is often a good idea.

```
[24]: # Time-series plot of mean PS differences between col_PS orig and col_PS lossy datasets,
      ↪grouped by month of year
ldcpy.plot(
    col_PS,
    "PS",
    sets=["orig", "lossy"],
    calc="mean",
    plot_type="time_series",
    group_by="time.month",
    calc_type="diff",
)
```



```
[25]: # Time-series plot of PS mean (grouped by month) in the original and lossy datasets
ldcpy.plot(
    col_PS,
    "PS",
    sets=["orig", "lossy"],
    calc="mean",
    plot_type="time_series",
    group_by="time.month",
)
```



```
[26]: del col_PS
```

## Surface Temperature

Now let's open the daily surface temperature (TS) data for 1920-2005 into a collection. These are larger files than the monthly PS data.

```
[27]: # load the first 86 years of daily surface temperature (TS) data
col_TS = ldcpy.open_datasets(
    "cam-fv",
    ["TS"],
    [
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/orig/TS.daily.19200101-20051231.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/lossy/c.TS.daily.19200101-20051231.nc"
    ],
    ["orig", "lossy"],
    chunks={"time": 500},
)
col_TS
```

dataset size in GB 13.89

```
[27]: <xarray.Dataset>
Dimensions:      (collection: 2, time: 31390, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 1920-01-01 00:00:00 ... 2005-12-31 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>,
↳ meta=np.ndarray>
    * collection (collection) <U5 'orig' 'lossy'
Data variables:
    TS          (collection, time, lat, lon) float32 dask.array<chunksize=(1, 500, 192,
↳ 288)>, meta=np.ndarray>
Attributes: (12/14)
  Conventions:      CF-1.0
  source:           CAM
  case:             b.e11.B20TRC5CNBDRD.f09_g16.031
  title:            UNSET
  logname:          mickelso
  host:             ys0219
  ...
  initial_file:     b.e11.B20TRC5CNBDRD.f09_g16.001.cam.i.1920-01-01-000000.nc
  topography_file:  /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:          Tue Nov 3 13:56:03 2020: ncks -L 5 TS.daily.19200101-2...
  NCO:              netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures:    area: cell_area
  data_type:        cam-fv
```

Look at the first time slice (time = 0) statistics:

```
[28]: ldcpy.compare_stats(col_TS.isel(time=0), "TS", ["orig", "lossy"])
```

```
<IPython.core.display.HTML object>
```

	orig	lossy
mean	284.49	284.43
variance	533.99	533.44
standard deviation	23.108	23.096
min value	216.73	216.69
max value	315.58	315.5
probability positive	1	1
number of zeros	0	0
spatial autocorr - latitude	0.99392	0.99392
spatial autocorr - longitude	0.9968	0.9968
entropy estimate	0.41487	0.13675

```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

	lossy
max abs diff	0.12497
min abs diff	0
mean abs diff	0.059427
mean squared diff	0.0035316
root mean squared diff	0.069462

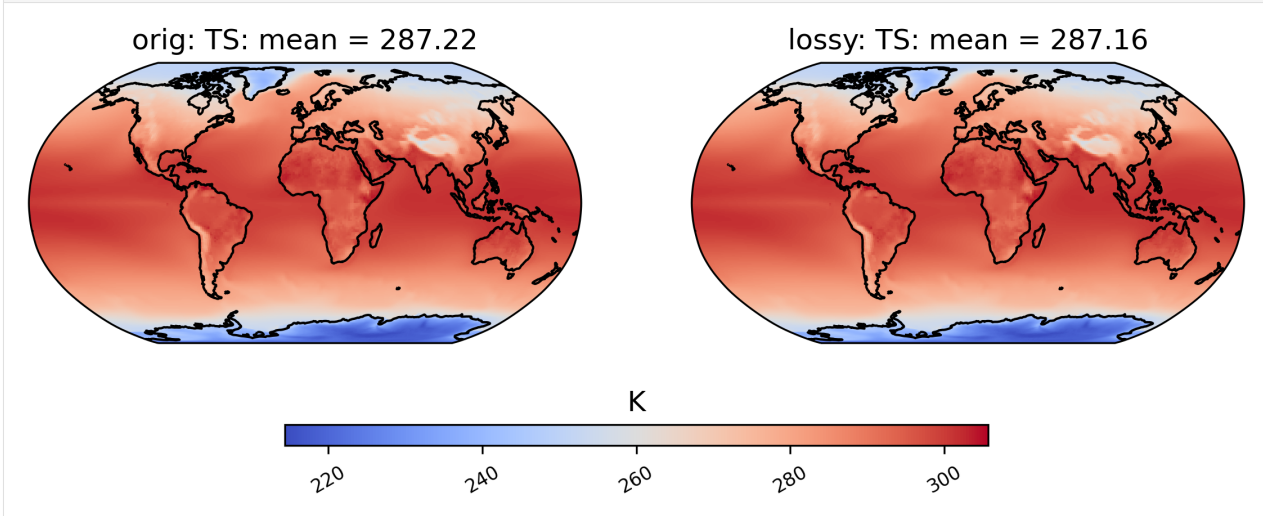
(continues on next page)

(continued from previous page)

normalized root mean squared diff	0.0006603
normalized max pointwise error	0.0012642
pearson correlation coefficient	1
ks p-value	0.36817
spatial relative error(% > 0.0001)	73.293
max spatial relative error	0.00048733
Data SSIM	0.97883

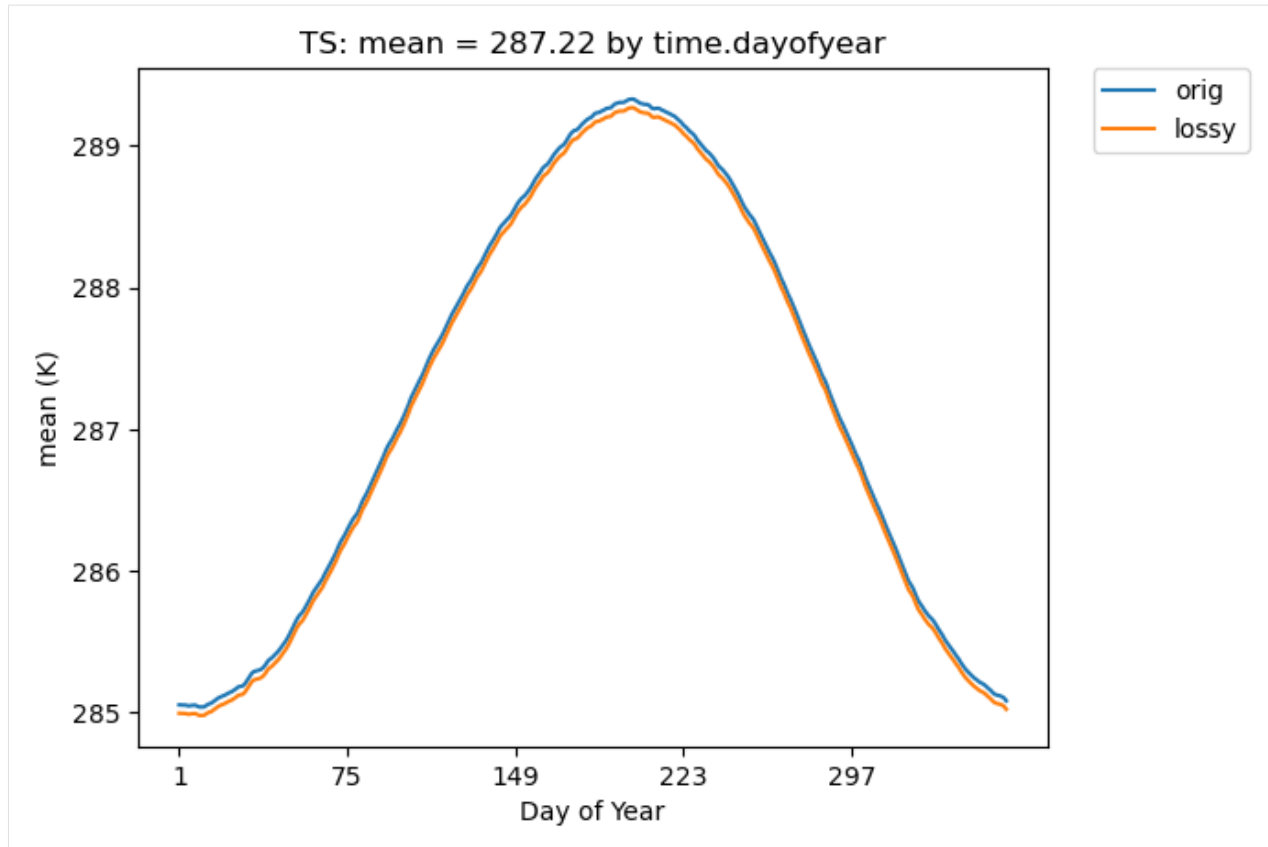
Now we compare mean TS over time in a plot:

```
[29]: # comparison between mean TS values in col_TS orig and lossy datasets
ldcpy.plot(col_TS, "TS", sets=["orig", "lossy"], calc="mean")
```



Below we do a time series plot and group by day of the year. (Note that the group\_by functionality is not fast.)

```
[30]: # Time-series plot of TS means (grouped by days) in the original and lossy datasets
ldcpy.plot(
    col_TS,
    "TS",
    sets=["orig", "lossy"],
    calc="mean",
    plot_type="time_series",
    group_by="time.dayofyear",
)
```



Let's delete the PS and TS data to free up memory.

```
[31]: del col_TS
```

## Precipitation Rate

Now let's open the daily precipitation rate (PRECT) data for 2006-2080 into a collection.

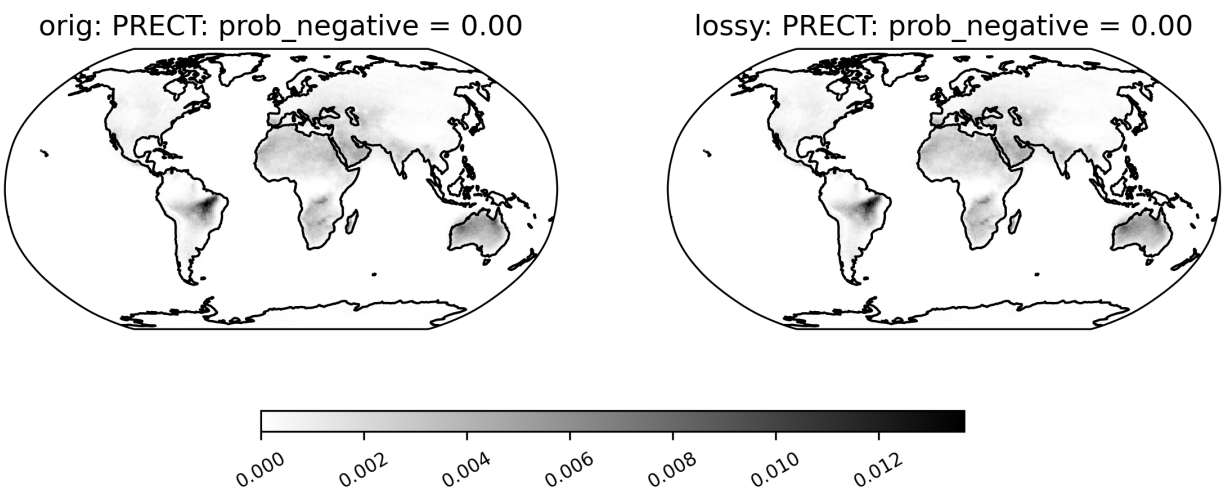
```
[32]: # load the last 75 years of PRECT data
col_PRECT = ldcpy.open_datasets(
    "cam-fv",
    ["PRECT"],
    [
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/orig/PRECT.daily.20060101-20801231.nc
    ↪",
        "/glade/p/cisl/asap/ldcpy_sample_data/lens/lossy/c.PRECT.daily.20060101-20801231.
    ↪nc",
    ],
    ["orig", "lossy"],
    chunks={"time": 500},
)
col_PRECT
```

dataset size in GB 12.11



```
[32]: <xarray.Dataset>
Dimensions:      (collection: 2, time: 27375, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 2006-01-01 00:00:00 ... 2080-12-31 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>,
↳ meta=np.ndarray>
  * collection   (collection) <U5 'orig' 'lossy'
Data variables:
  PRECT         (collection, time, lat, lon) float32 dask.array<chunksize=(1, 500, 192,
↳ 288)>, meta=np.ndarray>
Attributes: (12/14)
  Conventions:   CF-1.0
  source:        CAM
  case:          b.e11.BRCP85C5CNBDRD.f09_g16.031
  title:         UNSET
  logname:       mickelso
  host:          ys1023
  ...           ...
  initial_file:  b.e11.B20TRC5CNBDRD.f09_g16.031.cam.i.2006-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:       Tue Nov 3 14:13:51 2020: ncks -L 5 PRECT.daily.2006010...
  NCO:          netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures: area: cell_area
  data_type:     cam-fv
```

```
[34]: # compare probability of negative rainfall (and get ssim)
ldcpy.plot(col_PRECT, "PRECT", sets=["orig", "lossy"], calc="prob_negative", color=
↳ "binary")
```



Mean PRECT over time...

```
[35]: # Time-series plot of PRECT mean in 'orig' dataset
ldcpy.plot(
  col_PRECT,
```

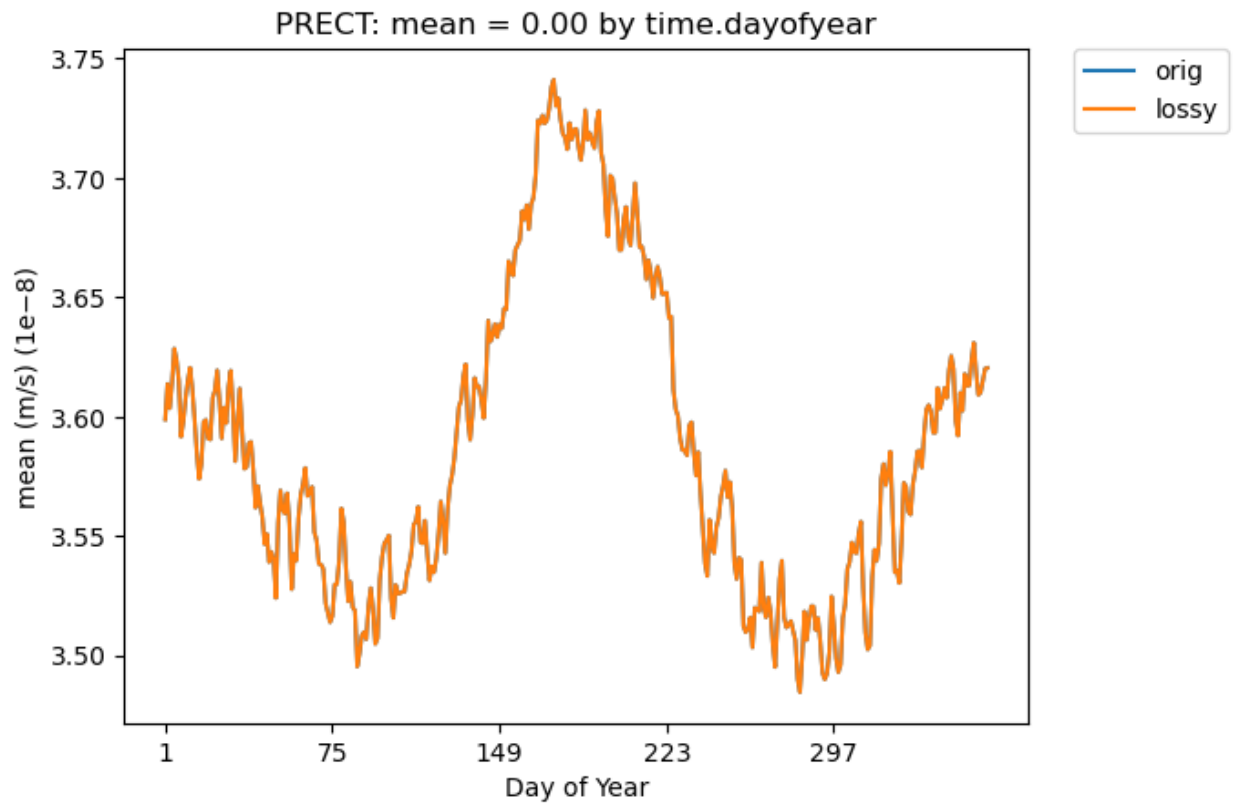
(continues on next page)

(continued from previous page)

```

"PRECT",
sets=["orig", "lossy"],
calc="mean",
plot_type="time_series",
group_by="time.dayofyear",
)

```

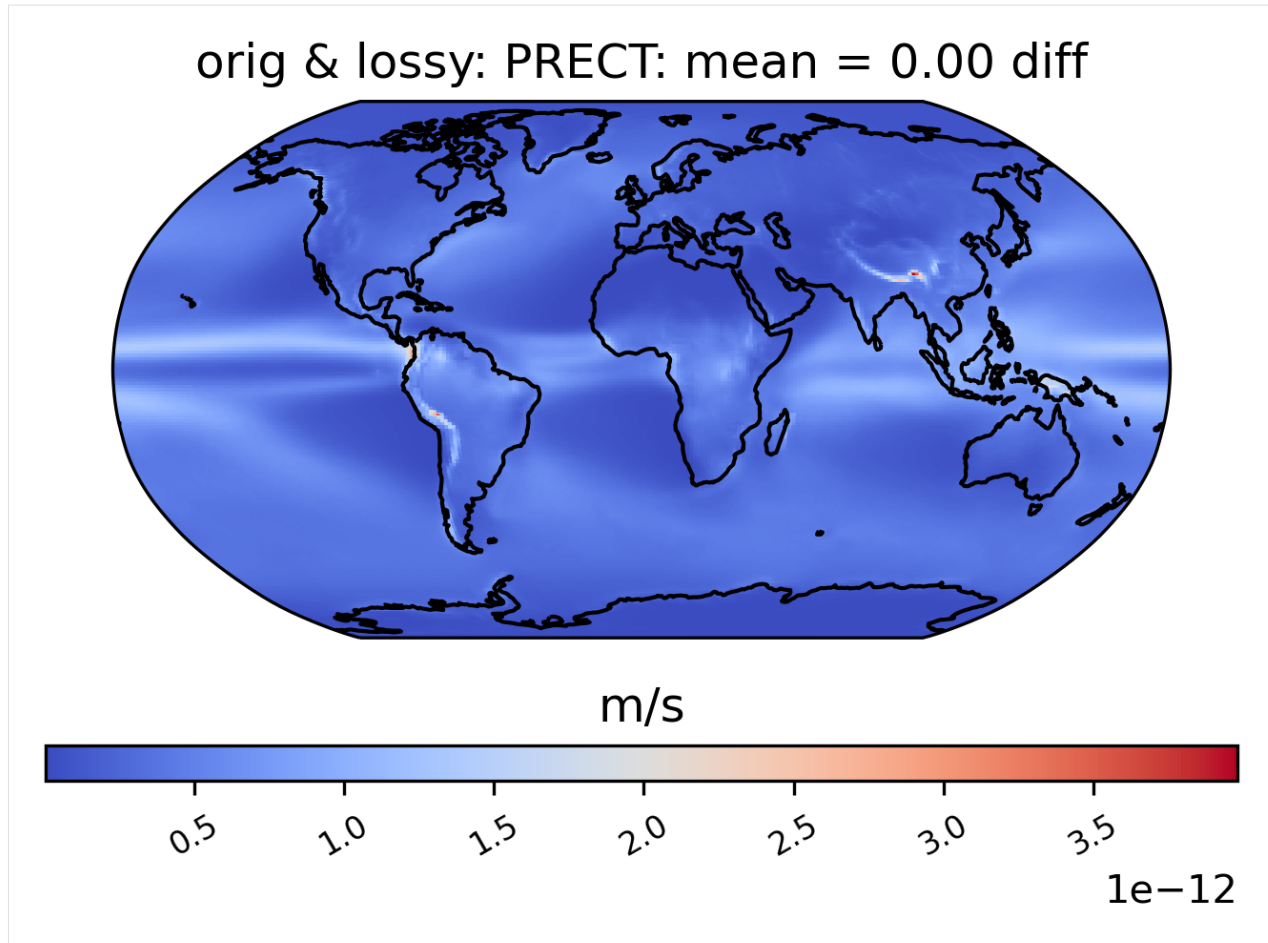


Now look at mean over time spatial plot:

```

[36]: # diff between mean PRECT values across the entire timeseries
ldcpy.plot(
    col_PRECT,
    "PRECT",
    sets=["orig", "lossy"],
    calc="mean",
    calc_type="diff",
)

```



Calculating the correlation of the lag-1 values ... for the first 10 years (This operation can be time-consuming).

```
[ ]: # plot of lag-1 correlation of PRECT values
ldcpy.plot(col_PRECT, "PRECT", sets=["orig", "lossy"], calc="lag1", start=0, end=3650)
```

## CAGEO Plots

Plots from different data sets used in CAGEO paper (Poppick et al., “A Statistical Analysis of Lossily Compressed Climate Model Data”, doi:10.1016/j.cageo.2020.104599) comparing the sz and zfp compressors with a number of metrics (specified with “calc” and “calc\_type” in the plot routine).

```
[37]: col_TS = ldcpy.open_datasets(
    "cam-fv",
    ["TS"],
    [
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/orig/b.e11.B20TRC5CNBDRD.f09_g16.030.
        ↪cam.h1.TS.19200101-20051231.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/lossy/sz1.0.TS.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/lossy/zfp1.0.TS.nc",
    ],
    ["orig", "sz1.0", "zfp1.0"],
```

(continues on next page)

(continued from previous page)

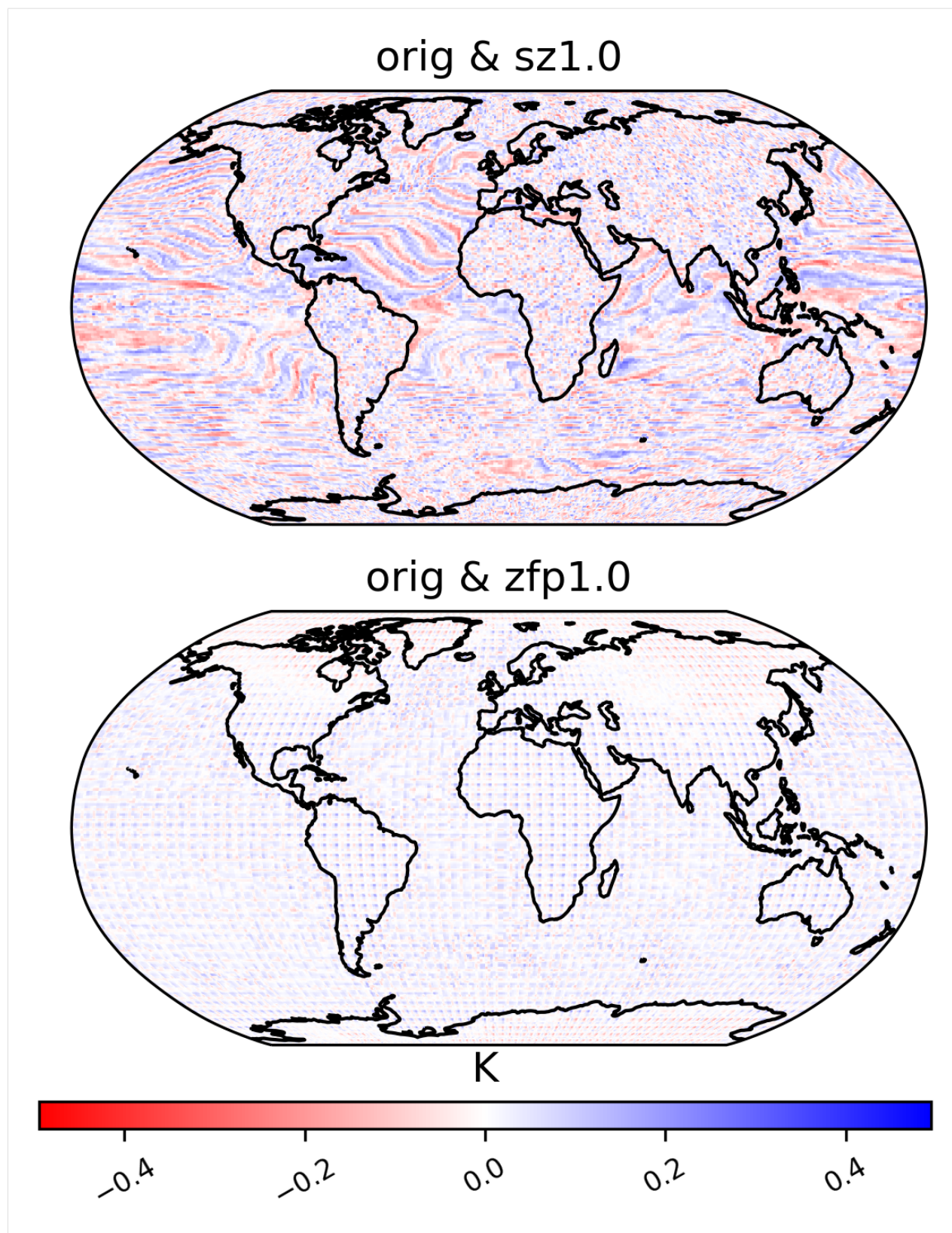
```
    chunks={"time": 500},
)
col_PRECT = ldcpy.open_datasets(
    "cam-fv",
    ["PRECT"],
    [
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/orig/b.e11.B20TRC5CNBDRD.f09_g16.030.
→cam.h1.PRECT.19200101-20051231.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/lossy/sz1e-8.PRECT.nc",
        "/glade/p/cisl/asap/ldcpy_sample_data/cageo/lossy/zfp1e-8.PRECT.nc",
    ],
    ["orig", "sz1e-8", "zfp1e-8"],
    chunks={"time": 500},
)
```

dataset size in GB 20.83

dataset size in GB 20.83

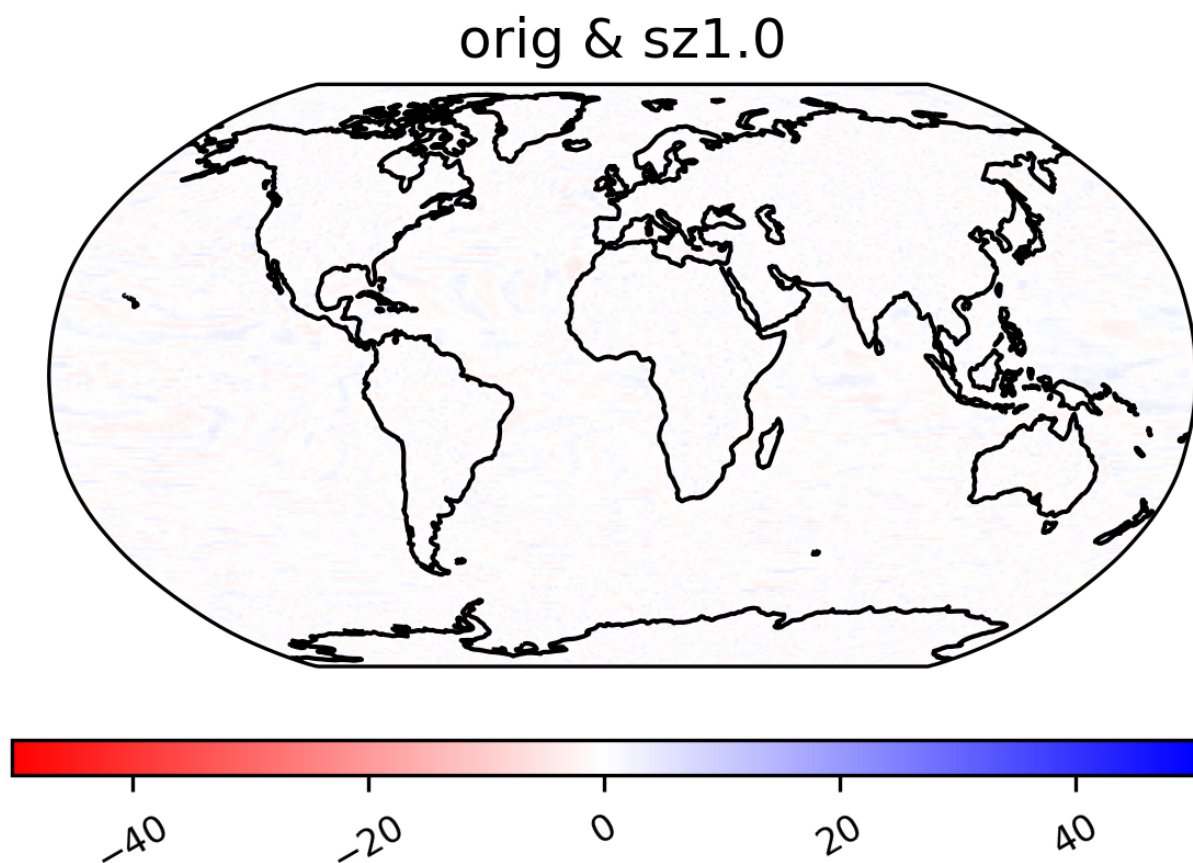
difference in mean

```
[38]: ldcpy.plot(
    col_TS,
    "TS",
    sets=["orig", "sz1.0", "zfp1.0"],
    calc="mean",
    calc_type="diff",
    color="bwr_r",
    start=0,
    end=50,
    tex_format=False,
    vert_plot=True,
    short_title=True,
    axes_symmetric=True,
)
```



Zscore

```
[39]: ldcpy.plot(  
    col_TS,  
    "TS",  
    sets=["orig", "sz1.0"],  
    calc="zscore",  
    calc_type="calc_of_diff",  
    color="bwr_r",  
    start=0,  
    end=1000,  
    tex_format=False,  
    vert_plot=True,  
    short_title=True,  
    axes_symmetric=True,  
)
```



Pooled variance ratio

```
[40]: ldcpy.plot(  
    col_TS,  
    "TS",  
    sets=["orig", "sz1.0"],  
    calc="pooled_var_ratio",  
    color="bwr_r",
```

(continues on next page)

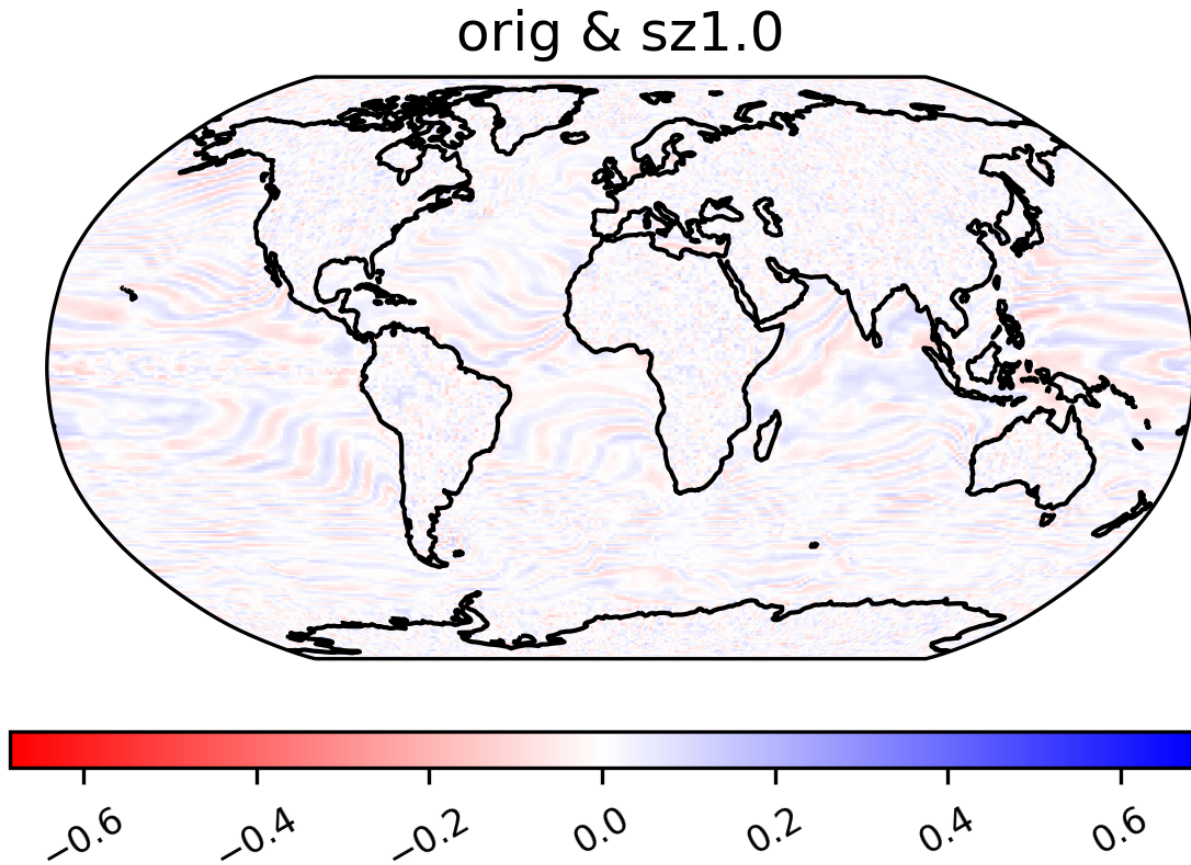


(continued from previous page)

```

calc_type="calc_of_diff",
transform="log",
start=0,
end=100,
tex_format=False,
vert_plot=True,
short_title=True,
axes_symmetric=True,
)

```



annual harmonic relative ratio

```

[41]: col_TS["TS"] = col_TS.TS.chunk({"time": -1, "lat": 10, "lon": 10})
ldcpy.plot(
    col_TS,
    "TS",
    sets=["orig", "sz1.0", "zfp1.0"],
    calc="ann_harmonic_ratio",
    transform="log",
    calc_type="calc_of_diff",
    tex_format=False,
    axes_symmetric=True,
)

```

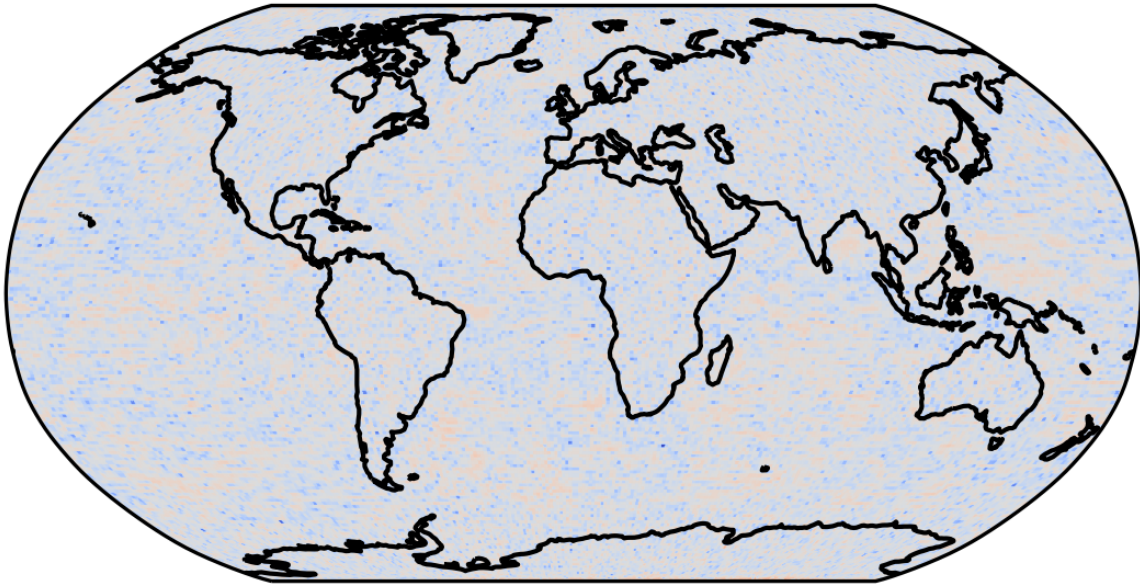
(continues on next page)

(continued from previous page)

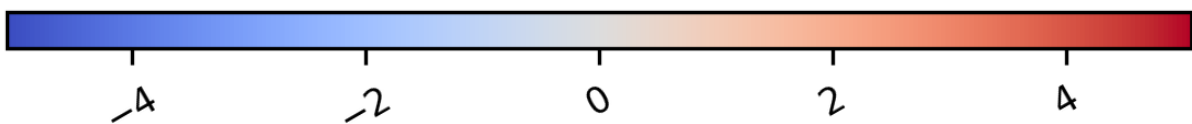
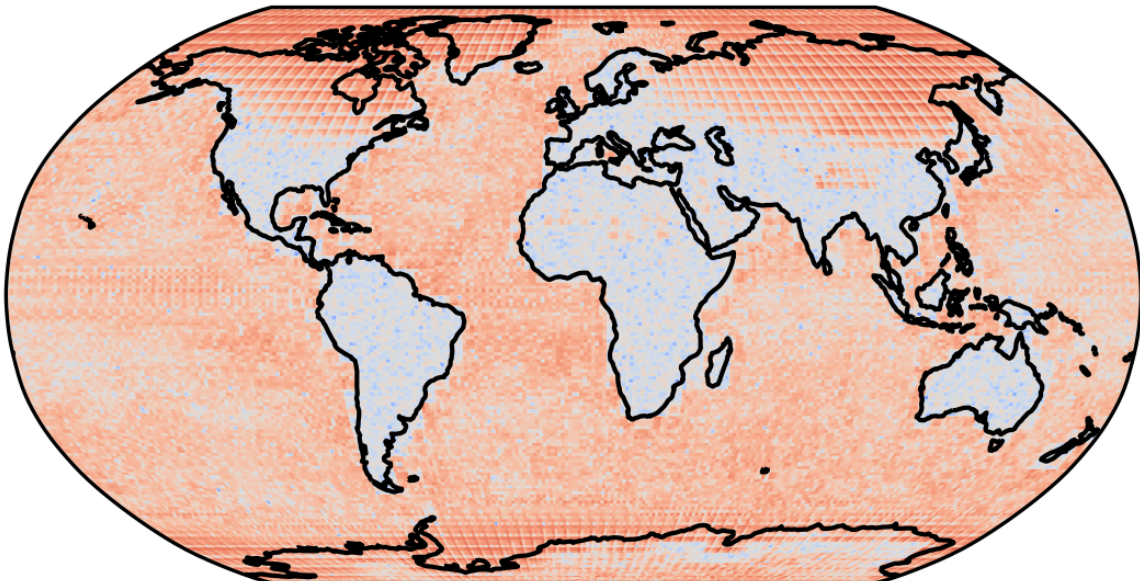
```
    vert_plot=True,  
    short_title=True,  
)  
col_TS["TS"] = col_TS.TS.chunk({"time": 500, "lat": 192, "lon": 288})
```



orig &amp; sz1.0

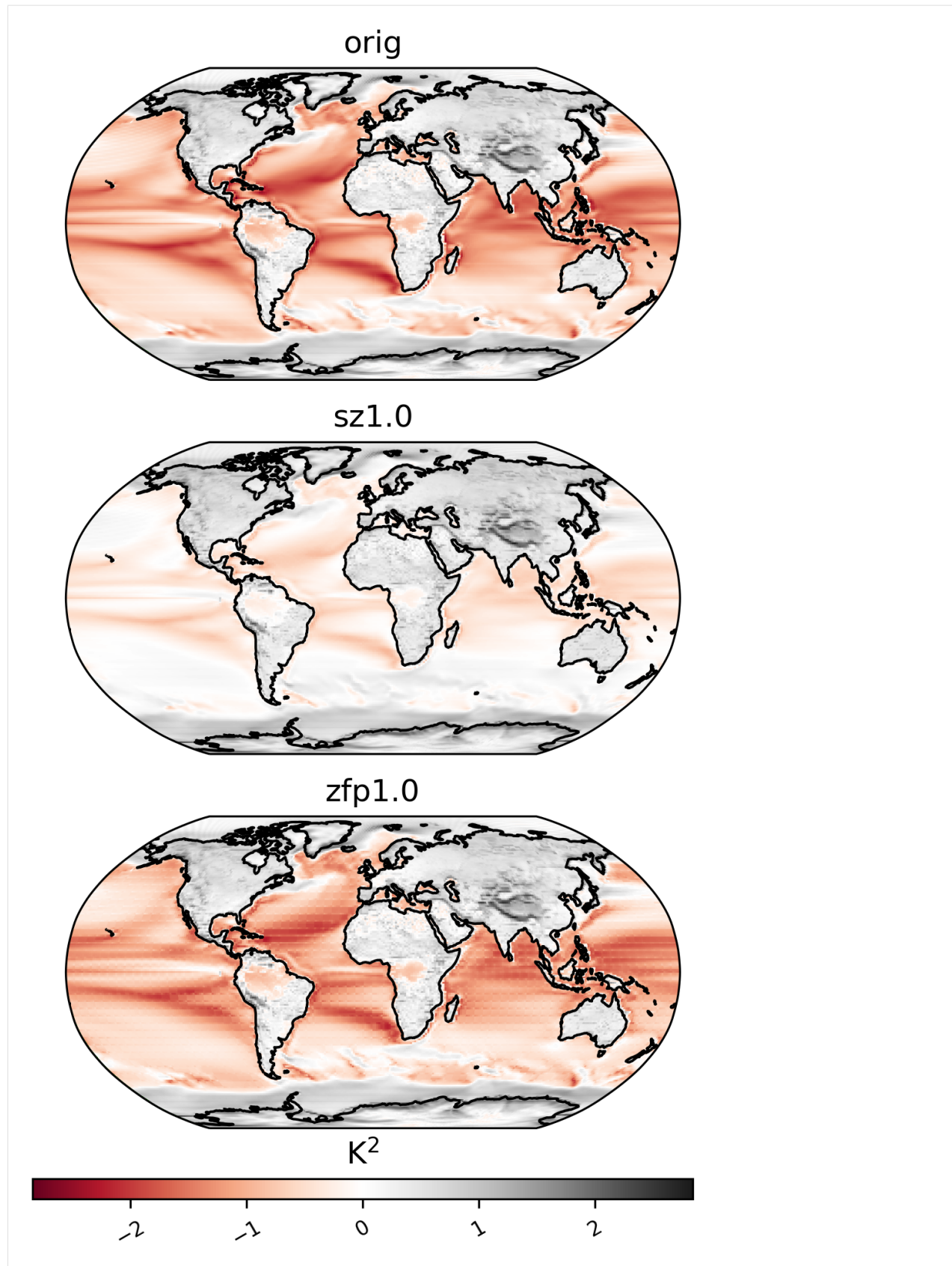


orig &amp; zfp1.0



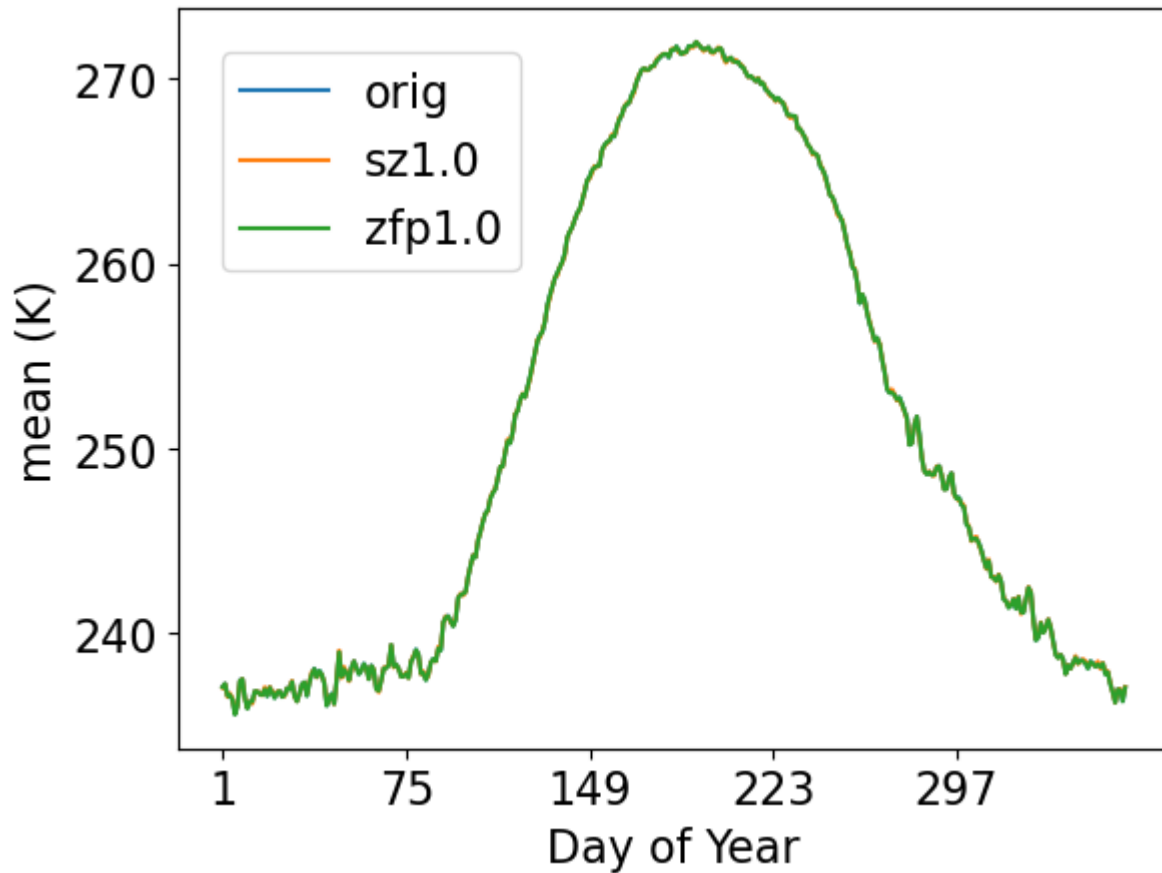
NS contrast variance

```
[42]: ldcpy.plot(  
    col_TS,  
    "TS",  
    sets=["orig", "sz1.0", "zfp1.0"],  
    calc="ns_con_var",  
    color="RdGy",  
    calc_type="raw",  
    transform="log",  
    axes_symmetric=True,  
    tex_format=False,  
    vert_plot=True,  
    short_title=True,  
)
```



mean by day of year

```
[43]: ldcpy.plot(
    col_TS,
    "TS",
    sets=["orig", "sz1.0", "zfp1.0"],
    calc="mean",
    plot_type="time_series",
    group_by="time.dayofyear",
    legend_loc="upper left",
    lat=90,
    lon=0,
    vert_plot=True,
    tex_format=False,
    short_title=True,
)
```



standardized mean errors by day of year

```
[44]: ldcpy.plot(
    col_TS,
    "TS",
    sets=["orig", "sz1.0", "zfp1.0"],
    calc="standardized_mean",
    legend_loc="lower left",
)
```

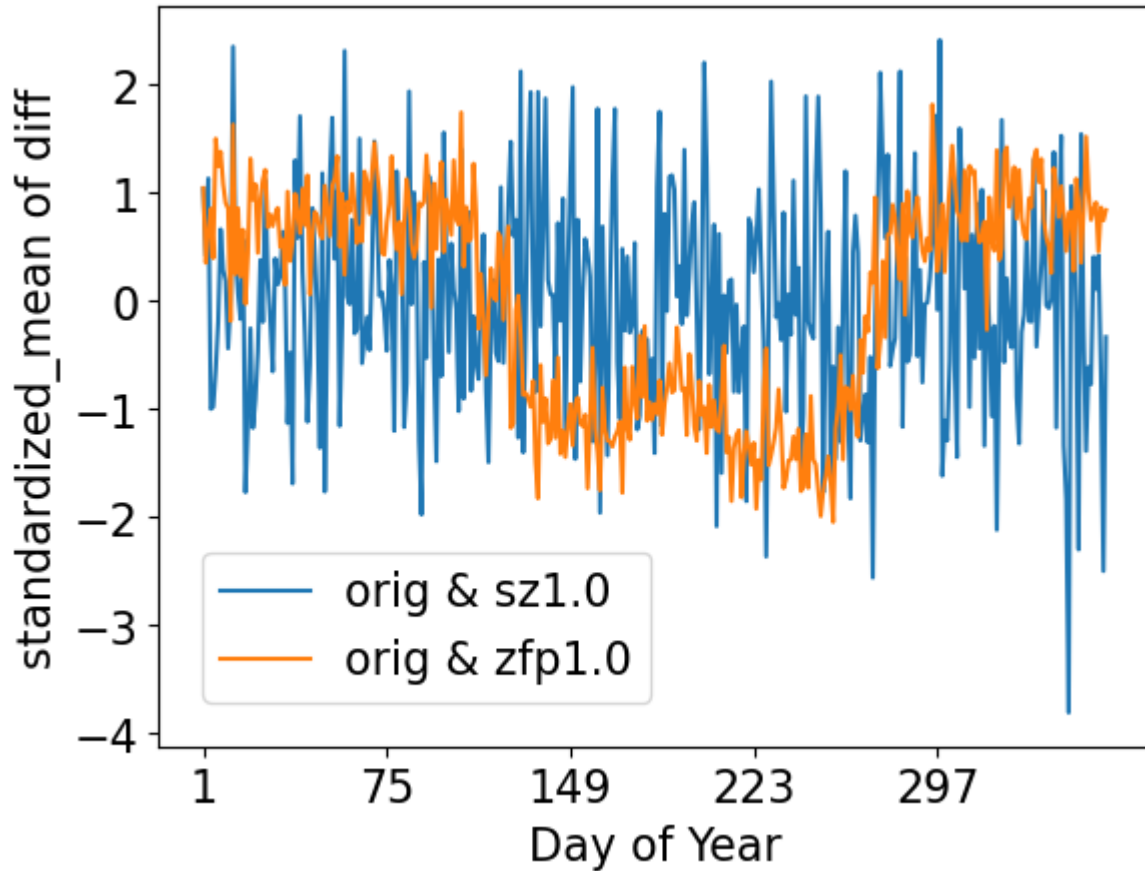
(continues on next page)

(continued from previous page)

```

calc_type="calc_of_diff",
plot_type="time_series",
group_by="time.dayofyear",
tex_format=False,
lat=90,
lon=0,
vert_plot=True,
short_title=True,
)

```



```
[45]: del col_TS
```

Do any other comparisons you wish ... and then clean up!

```
[46]: client.close()
```

```
[ ]:
```

This Notebook demonstrates comparing different levels of compression on several test variables in the CESM-LENS1 dataset (<http://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html>). In doing so, we will start a DASK client from Jupyter.

## 2.4.6 Setup

Before using this notebook, we recommend familiarity with the TutorialNotebook and the LargeDataGladeNotebook. This notebook is meant to be run on NCAR's JupyterHub (<https://jupyterhub.hpc.ucar.edu>). The subset of CESM-LENS1 data on glade is located in `/glade/p/cisl/asap/abaker/compression_samples/cam-lens`.

When you launch your NCAR JupyterHub session, you will need to indicate a machine (Cheyenne or Casper) and then you will need your charge account. You can then launch the session and navigate to this notebook.

NCAR's JupyterHub documentation: <https://www2.cisl.ucar.edu/resources/jupyterhub-ncar>

**You need to run your notebook with the “cmip6-201910” kernel (choose from the dropdown in the upper left.)**

When you launch your NCAR JupyterHub session, you will need to indicate a machine and then you will need your charge account. You can then launch the session and navigate to this notebook.

NCAR's JupyterHub documentation: <https://www2.cisl.ucar.edu/resources/jupyterhub-ncar>

```
[1]: # Make sure you are using the cmip6-2019.10 kernel

# Add ldcpy root to system path (MODIFY FOR YOUR LDCPY CODE LOCATION)
import sys

sys.path.insert(0, '/glade/u/home/abaker/repos/ldcpy')
import ldcpy

# Display output of plots directly in Notebook
%matplotlib inline
# Automatically reload module if it is edited
%reload_ext autoreload
%autoreload 2

# silence warnings
import warnings

warnings.filterwarnings("ignore")
```

Start DASK... and connect to client

```
[2]: from dask.distributed import Client
from ncar_jobqueue import NCARCluster

cluster = NCARCluster(project='NTDD0004')
# scale as needed
cluster.adapt(minimum_jobs=1, maximum_jobs=30)
cluster
client = Client(cluster)
```



## 2.4.7 Sample Data

In the glade directory mentionned above (/glade/p/cisl/asap/abaker/compression\_samples/cam-lens), we have an “orig” directory with the original (uncompressed) version of each sample variable: Q, FLNS, TREFHT, TMQ, PS, U, PRECT, LHFLX, CCN3, TS, and CLOUD. Then there is a directory for each sample variable with a variety of compressed variants.

Note that each test variables has a two timeseries slices: 1920-2005 and 2006-2080. The range of data is indicated in the filename along with whether the data is “daily” or “monthly” timeslices.

Variables U, CLOUD, and Q are 3D, and the rest are 2D.

```
[3]: # list directory contents
import os
```

```
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens")
```

```
[3]: ['README.txt',
      'Q',
      'FLNS',
      'TREFHT',
      'TMQ',
      'PS',
      'U',
      'PRECT',
      'LHFLX',
      'CCN3',
      'orig',
      'CLOUD',
      'TS']
```

```
[4]: # List the original data files
```

```
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens/orig")
```

```
[4]: ['CLOUD.monthly.200601-208012.nc',
      'LHFLX.daily.20060101-20801231.nc',
      'TREFHT.monthly.192001-200512.nc',
      'U.monthly.192001-200512.nc',
      'PRECT.daily.20060101-20801231.nc',
      'CLOUD.monthly.192001-200512.nc',
      'TMQ.monthly.192001-200512.nc',
      'CCN3.monthly.200601-208012.nc',
      'PS.monthly.192001-200512.nc',
      'CCN3.monthly.192001-200512.nc',
      'PRECT.daily.19200101-20051231.nc',
      'TS.daily.20060101-20801231.nc',
      'PS.monthly.200601-208012.nc',
      'TS.daily.19200101-20051231.nc',
      'Q.monthly.192001-200512.nc',
      'Q.monthly.200601-208012.nc',
      'TMQ.monthly.200601-208012.nc',
      'U.monthly.200601-208012.nc',
      'TREFHT.monthly.200601-208012.nc',
      'FLNS.monthly.192001-200512.nc',
      'FLNS.monthly.200601-208012.nc',
```

(continues on next page)

(continued from previous page)

```
'LHFLX.daily.19200101-20051231.nc']
```

```
[5]: # List the compressed data files for TMQ
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ")
```

```
[5]: ['zfp.p22.TMQ.monthly.192001-200512.nc',
      'zfp.p18.TMQ.monthly.200601-208012.nc',
      'zfp.p22.TMQ.monthly.200601-208012.nc',
      'zfp.p20.TMQ.monthly.192001-200512.nc',
      'zfp.p12.TMQ.monthly.200601-208012.nc',
      'zfp.p8.TMQ.monthly.192001-200512.nc',
      'zfp.p18.TMQ.monthly.192001-200512.nc',
      'zfp.p12.TMQ.monthly.192001-200512.nc',
      'zfp.p8.TMQ.monthly.200601-208012.nc',
      'zfp.p10.TMQ.monthly.192001-200512.nc',
      'fpzip16.TMQ.monthly.200601-208012.nc',
      'zfp.p16.TMQ.monthly.192001-200512.nc',
      'zfp.p20.TMQ.monthly.200601-208012.nc',
      'zfp.p10.TMQ.monthly.200601-208012.nc',
      'zfp.p14.TMQ.monthly.200601-208012.nc',
      'zfp.p16.TMQ.monthly.200601-208012.nc',
      'zfp.p14.TMQ.monthly.192001-200512.nc',
      'fpzip16.TMQ.monthly.192001-200512.nc']
```

Make a collection with some of the TMQ data to compare. We'll look at the 2006-2080 data. Make sure to use “useful” labels. You can't mix the 2006-2080 and 1920-2005 data in the same collection as they have different numbers of time slices.

The “fpzip” version is from the blind evaluation. For the “zfp” versions, the p indicates the precision parameter (the higher the number after p, the more accurate).

```
[6]: col_tmq = ldcpy.open_datasets(
    "cam-fv",
    ["TMQ"],
    [
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/orig/TMQ.monthly.200601-
        ↪208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/fpzip16.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p8.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p10.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p12.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p14.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p16.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p18.TMQ.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p20.TMQ.monthly.
        ↪200601-208012.nc",
    ]
)
```

(continues on next page)



(continued from previous page)

```

        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TMQ/zfp.p22.TMQ.monthly.
↪200601-208012.nc",
    ],
    [
        "orig",
        "fpzip16",
        "zfp-p8",
        "zfp-p10",
        "zfp-p12",
        "zfp-p14",
        "zfp-p16",
        "zfp-p18",
        "zfp-p20",
        "zfp-p22",
    ],
    chunks={"time": 700},
)
col_tmq

```

dataset size in GB 2.00

```

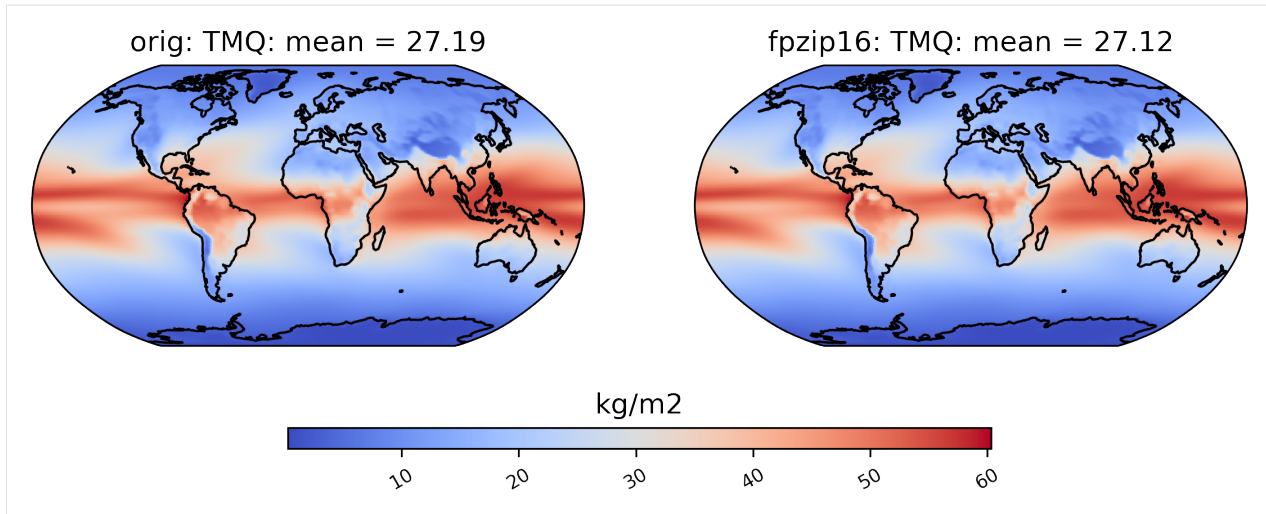
[6]: <xarray.Dataset>
Dimensions:      (collection: 10, time: 900, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 2006-02-01 00:00:00 ... 2081-01-01 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>
↪meta=np.ndarray>
  * collection   (collection) <U7 'orig' 'fpzip16' ... 'zfp-p20' 'zfp-p22'
Data variables:
  TMQ            (collection, time, lat, lon) float32 dask.array<chunksize=(1, 700, 192,
↪288)>, meta=np.ndarray>
Attributes: (12/14)
  Conventions:   CF-1.0
  source:        CAM
  case:          b.e11.BRCP85C5CNBDRD.f09_g16.031
  title:         UNSET
  logname:       mickelso
  host:          ys1023
  ...           ...
  initial_file:  b.e11.B20TRC5CNBDRD.f09_g16.031.cam.i.2006-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:       Tue Nov 3 14:06:43 2020: ncks -L 5 TMQ.monthly.200601-...
  NCO:           netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures: area: cell_area
  data_type:     cam-fv

```

```

[7]: # first compare the mean TMQ values in col_tmq orig and fpzip datasets (which is in the
↪released lens data)
ldcpy.plot(col_tmq, "TMQ", sets=["orig", "fpzip16"], calc="mean")

```



Let's look at some statistics for the time slice= 100 of the data (the data has 900 time slices)

```
[8]: ldcpy.compare_stats(col_tmq.isel(time=100), "TMQ", ["orig", "fpzip16"])
```

<IPython.core.display.HTML object>

	orig	fpzip16
mean	25.015	24.945
variance	220.84	219.64
standard deviation	14.861	14.82
min value	0.16246	0.16211
max value	56.345	56.25
probability positive	1	1
number of zeros	0	0
spatial autocorr - latitude	0.99717	0.99717
spatial autocorr - longitude	0.99802	0.99802
entropy estimate	0.48197	0.13009

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

	fpzip16
max abs diff	0.24997
min abs diff	0
mean abs diff	0.069938
mean squared diff	0.0048913
root mean squared diff	0.093516
normalized root mean squared diff	0.0013828
normalized max pointwise error	0.0044493
pearson correlation coefficient	1
ks p-value	0.0067025
spatial relative error(% > 0.0001)	98.168
max spatial relative error	0.0077457
Data SSIM	0.99043

```
[9]: # The compression ratio (CR) for fpzip was ~3.4x better than lossless
# zfp - p10 has a similar CR - look at it's stats and also zfp-p12 (CR ~2.5x)
ldcpy.compare_stats(col_tmq.isel(time=100), "TMQ", ["orig", "zfp-p10", "zfp-p12"])
```

```
<IPython.core.display.HTML object>
```

	orig	zfp-p10	zfp-p12
mean	25.015	25.04	25.021
variance	220.84	221.26	220.94
standard deviation	14.861	14.875	14.864
min value	0.16246	0.16162	0.16248
max value	56.345	56.5	56.281
probability positive	1	1	1
number of zeros	0	0	0
spatial autocorr - latitude	0.99717	0.99715	0.99717
spatial autocorr - longitude	0.99802	0.998	0.99802
entropy estimate	0.48197	0.15017	0.22747

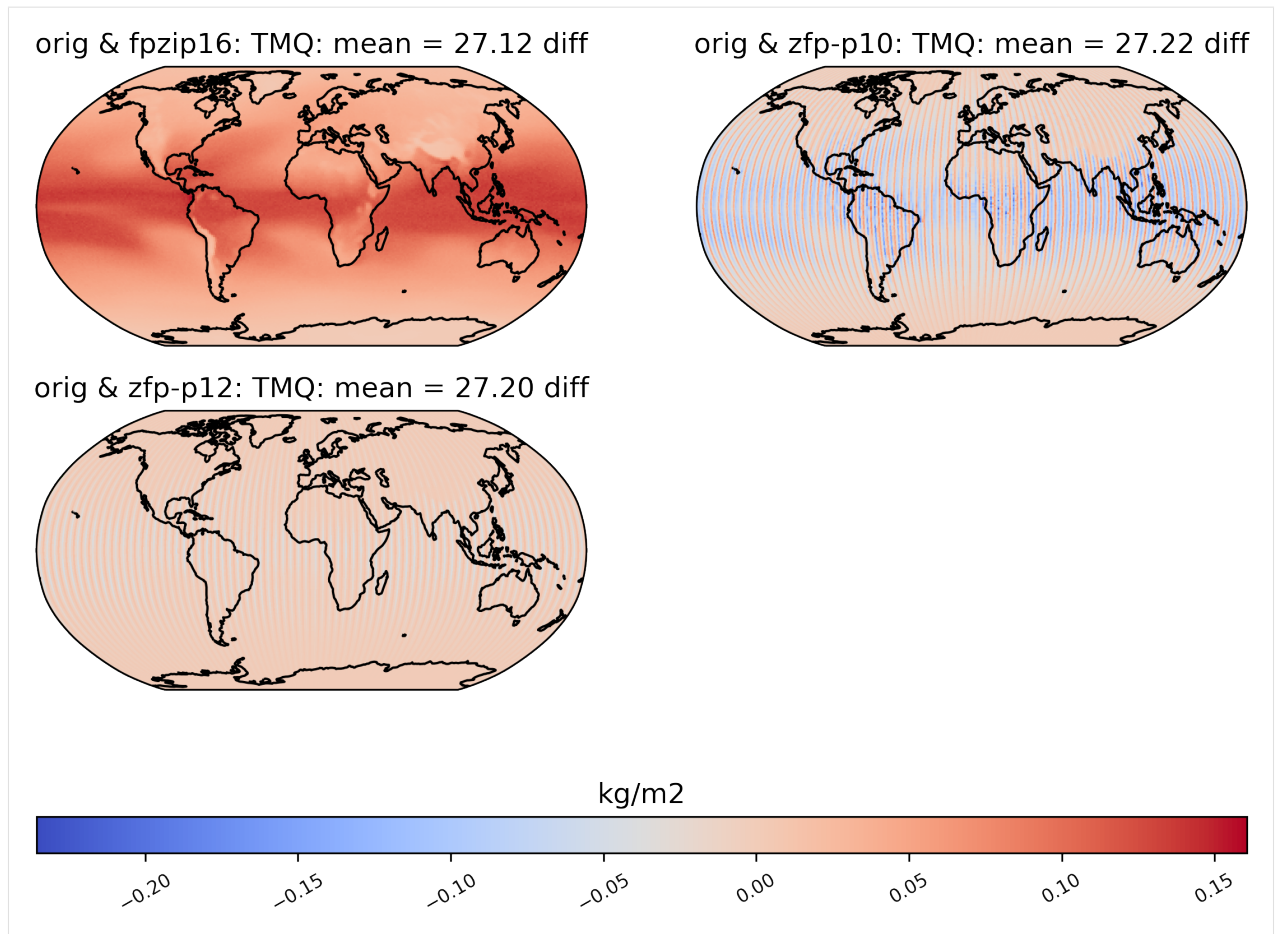
```
<IPython.core.display.HTML object>
```

```
<IPython.core.display.HTML object>
```

	zfp-p10	zfp-p12
max abs diff	0.53075	0.13838
min abs diff	2.5332e-07	0
mean abs diff	0.073004	0.019602
mean squared diff	0.00058902	3.5239e-05
root mean squared diff	0.10428	0.028061
normalized root mean squared diff	0.0015421	0.00041483
normalized max pointwise error	0.0082268	0.0020934
pearson correlation coefficient	0.99998	1
ks p-value	0.13637	0.36025
spatial relative error(% > 0.0001)	97.79	91.296
max spatial relative error	0.030925	0.0046611
Data SSIM	0.9872	0.99688

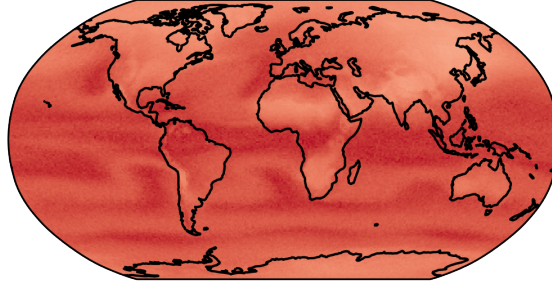
Let's look at something more interesting...

```
[10]: # diff between mean TS values in col_ds "orig" and "zfpA1.0" datasets
ldcpy.plot(
    col_tmq,
    "TMQ",
    sets=["orig", "fpzip16", "zfp-p10", "zfp-p12"],
    calc="mean",
    calc_type="diff",
)
```

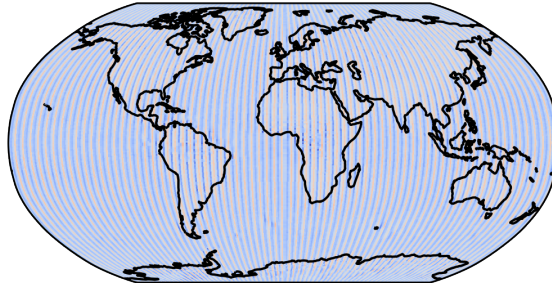


```
[11]: # plot of z-score under null hypothesis that orig value= compressed value
ldcpy.plot(
    col_tmq,
    "TMQ",
    sets=["orig", "fpzip16", "zfp-p10", "zfp-p12"],
    calc="zscore",
    calc_type="calc_of_diff",
)
```

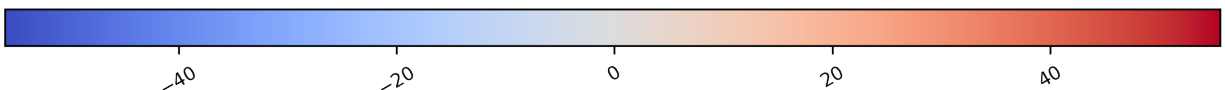
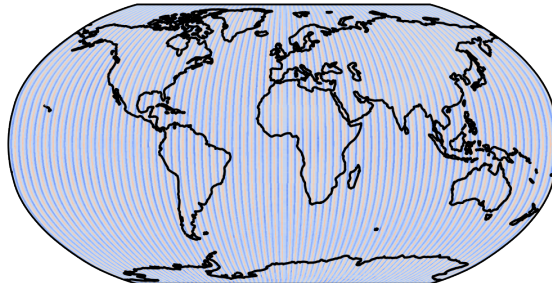
orig & fpzip16: TMQ: zscore: cutoff inf, % sig: 100.00 calc\_of\_diff



orig & zfp-p10: TMQ: zscore: cutoff inf, % sig: 79.52 calc\_of\_diff

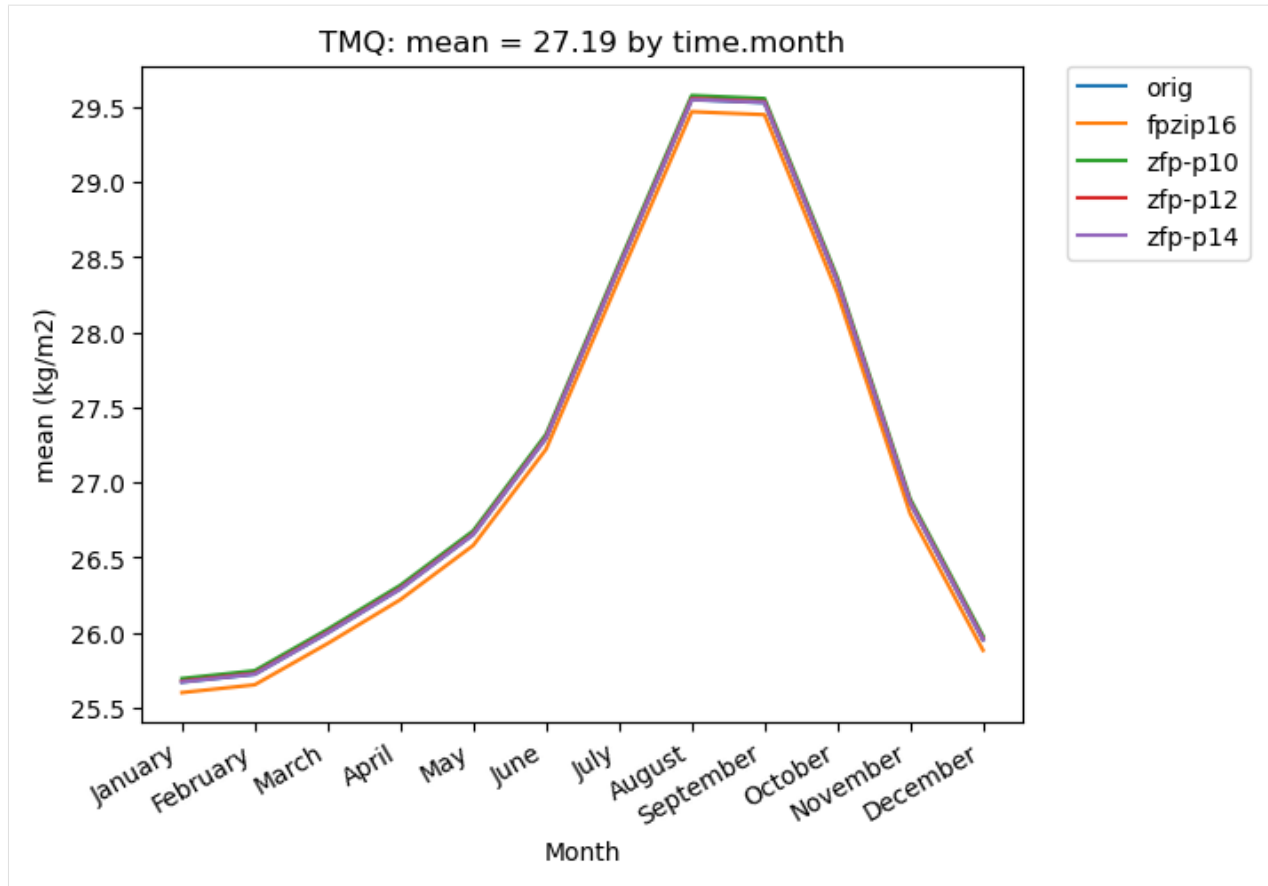


orig & zfp-p12: TMQ: zscore: cutoff inf, % sig: 73.63 calc\_of\_diff



[12]: *# Time-series plot of TMQ mean in the original and lossy datasets*

```
ldcpy.plot(
    col_tmq,
    "TMQ",
    sets=["orig", "fpzip16", "zfp-p10", "zfp-p12", "zfp-p14"],
    calc="mean",
    plot_type="time_series",
    group_by="time.month",
)
```



Now we look at other variables below

```
[13]: # List the compressed data files for TREFHT
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT")
```

```
[13]: ['zfp.p22.TREFHT.monthly.200601-208012.nc',
'zfp.p16.TREFHT.monthly.200601-208012.nc',
'zfp.p14.TREFHT.monthly.200601-208012.nc',
'zfp.p16.TREFHT.monthly.192001-200512.nc',
'zfp.p10.TREFHT.monthly.192001-200512.nc',
'zfp.p14.TREFHT.monthly.192001-200512.nc',
'zfp.p20.TREFHT.monthly.200601-208012.nc',
'zfp.p10.TREFHT.monthly.200601-208012.nc',
'zfp.p18.TREFHT.monthly.192001-200512.nc',
'zfp.p8.TREFHT.monthly.192001-200512.nc',
'zfp.p22.TREFHT.monthly.192001-200512.nc',
'fpzip20.TREFHT.monthly.200601-208012.nc',
'zfp.p12.TREFHT.monthly.200601-208012.nc',
'zfp.p20.TREFHT.monthly.192001-200512.nc',
'zfp.p12.TREFHT.monthly.192001-200512.nc',
'zfp.p18.TREFHT.monthly.200601-208012.nc',
'fpzip20.TREFHT.monthly.192001-200512.nc',
'zfp.p8.TREFHT.monthly.200601-208012.nc']
```

```
[14]: col_trefht = ldcpy.open_datasets(
    "cam-fv",
    ["TREFHT"],
    [
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/orig/TREFHT.monthly.
        ↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/fpzip20.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p8.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p10.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p12.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p14.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p16.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p18.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p20.TREFHT.
        ↪monthly.200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/TREFHT/zfp.p22.TREFHT.
        ↪monthly.200601-208012.nc",
    ],
    [
        "orig",
        "fpzip20",
        "zfp-p8",
        "zfp-p10",
        "zfp-p12",
        "zfp-p14",
        "zfp-p16",
        "zfp-p18",
        "zfp-p20",
        "zfp-p22",
    ],
    chunks={"time": 700},
)
col_trefht
```

dataset size in GB 2.00

```
[14]: <xarray.Dataset>
Dimensions:      (collection: 10, time: 900, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 2006-02-01 00:00:00 ... 2081-01-01 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunks=(192, 1, 288)>,
    ↪meta=np.ndarray>
  * collection   (collection) <U7 'orig' 'fpzip20' ... 'zfp-p20' 'zfp-p22'
```

Data variables:

(continues on next page)

(continued from previous page)

```

TREFHT      (collection, time, lat, lon) float32 dask.array<chunksize=(1, 700, 192, 288), meta=np.ndarray>
Attributes: (12/14)
  Conventions: CF-1.0
  source:      CAM
  case:        b.e11.BRCP85C5CNBDRD.f09_g16.031
  title:       UNSET
  logname:     mickelso
  host:        ys1023
  ...
  initial_file: b.e11.B20TRC5CNBDRD.f09_g16.031.cam.i.2006-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:      Tue Nov 3 14:11:24 2020: ncks -L 5 TREFHT.monthly.2006...
  NCO:          netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures: area: cell_area
  data_type:    cam-fv

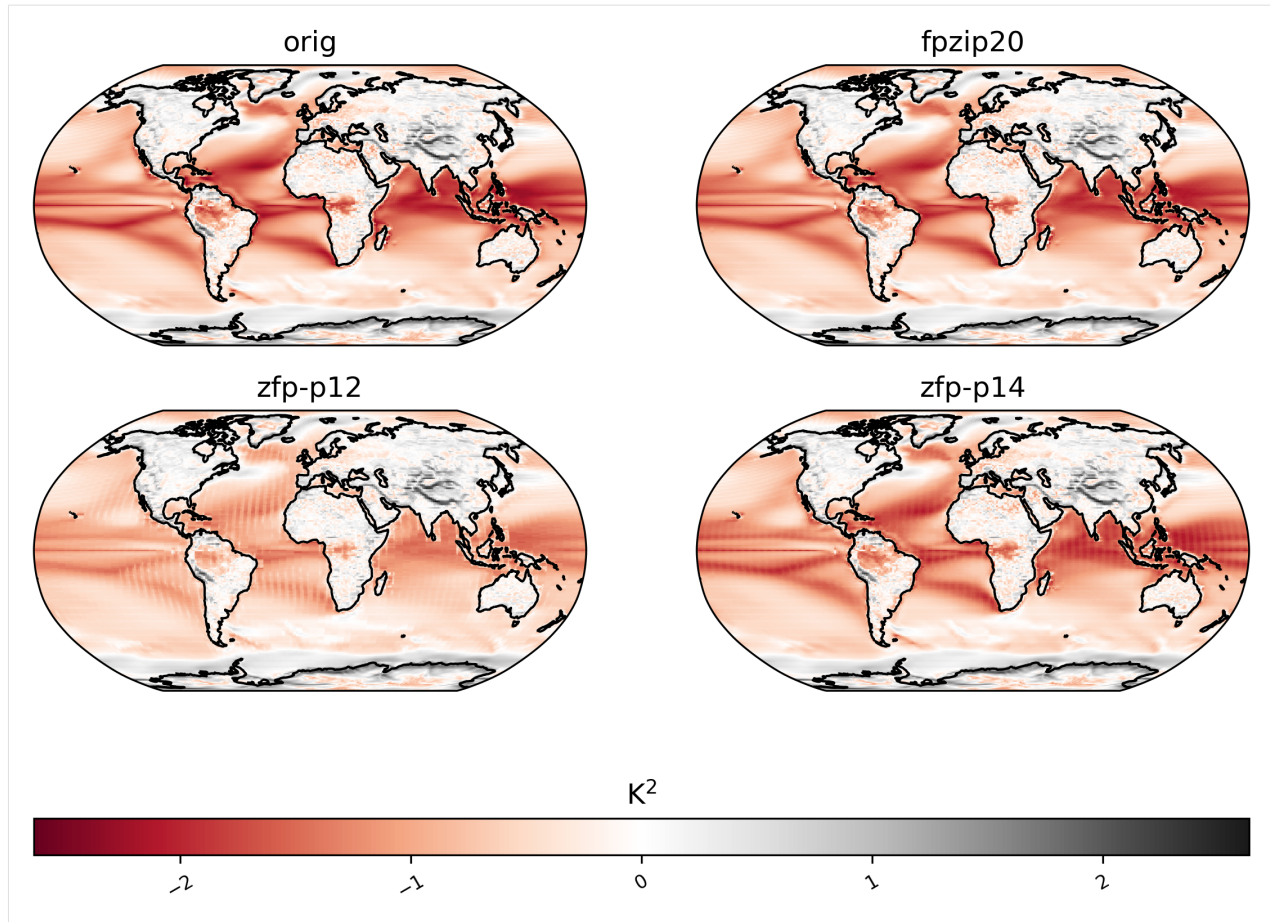
```

```

[15]: # Let's look and compare the north south contrast variances
ldcpy.plot(
    col_trefht,
    "TREFHT",
    sets=["orig", "fpzip20", "zfp-p12", "zfp-p14"],
    calc="ns_con_var",
    color="RdGy",
    calc_type="raw",
    transform="log",
    axes_symmetric=True,
    short_title=True,
)

```





```
[16]: # List the compressed data files for LHFLX
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX")
```

```
[16]: ['zfp.p14.LHFLX.daily.19200101-20051231.nc',
      'zfp.p12.LHFLX.daily.19200101-20051231.nc',
      'zfp.p18.LHFLX.daily.20060101-20801231.nc',
      'zfp.p12.LHFLX.daily.20060101-20801231.nc',
      'zfp.p16.LHFLX.daily.20060101-20801231.nc',
      'zfp.p22.LHFLX.daily.20060101-20801231.nc',
      'fpzip16.LHFLX.daily.19200101-20051231.nc',
      'zfp.p10.LHFLX.daily.19200101-20051231.nc',
      'fpzip16.LHFLX.daily.20060101-20801231.nc',
      'zfp.p8.LHFLX.daily.19200101-20051231.nc',
      'zfp.p8.LHFLX.daily.20060101-20801231.nc',
      'zfp.p20.LHFLX.daily.20060101-20801231.nc',
      'zfp.p14.LHFLX.daily.20060101-20801231.nc',
      'zfp.p10.LHFLX.daily.20060101-20801231.nc',
      'zfp.p16.LHFLX.daily.19200101-20051231.nc',
      'zfp.p22.LHFLX.daily.19200101-20051231.nc',
      'zfp.p18.LHFLX.daily.19200101-20051231.nc',
      'zfp.p20.LHFLX.daily.19200101-20051231.nc']
```

```
[17]: col_lhflx = ldcpy.open_datasets(
```

(continues on next page)

(continued from previous page)

```

    "cam-fv",
    ["LHFLX"],
    [
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/orig/LHFLX.daily.
↪20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/fpzip16.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p8.LHFLX.daily.
↪20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p10.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p12.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p14.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p16.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p18.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p20.LHFLX.
↪daily.20060101-20801231.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/LHFLX/zfp.p22.LHFLX.
↪daily.20060101-20801231.nc",
    ],
    [
        "orig",
        "fpzip16",
        "zfp-p8",
        "zfp-p10",
        "zfp-p12",
        "zfp-p14",
        "zfp-p16",
        "zfp-p18",
        "zfp-p20",
        "zfp-p22",
    ],
    chunks={"time": 100},
)
col_lhflx

```

dataset size in GB 60.55

```

[17]: <xarray.Dataset>
Dimensions:      (collection: 10, time: 27375, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 2006-01-01 00:00:00 ... 2080-12-31 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>,
↪meta=np.ndarray>
  * collection   (collection) <U7 'orig' 'fpzip16' ... 'zfp-p20' 'zfp-p22'
Data variables:

```

(continues on next page)

(continued from previous page)

```

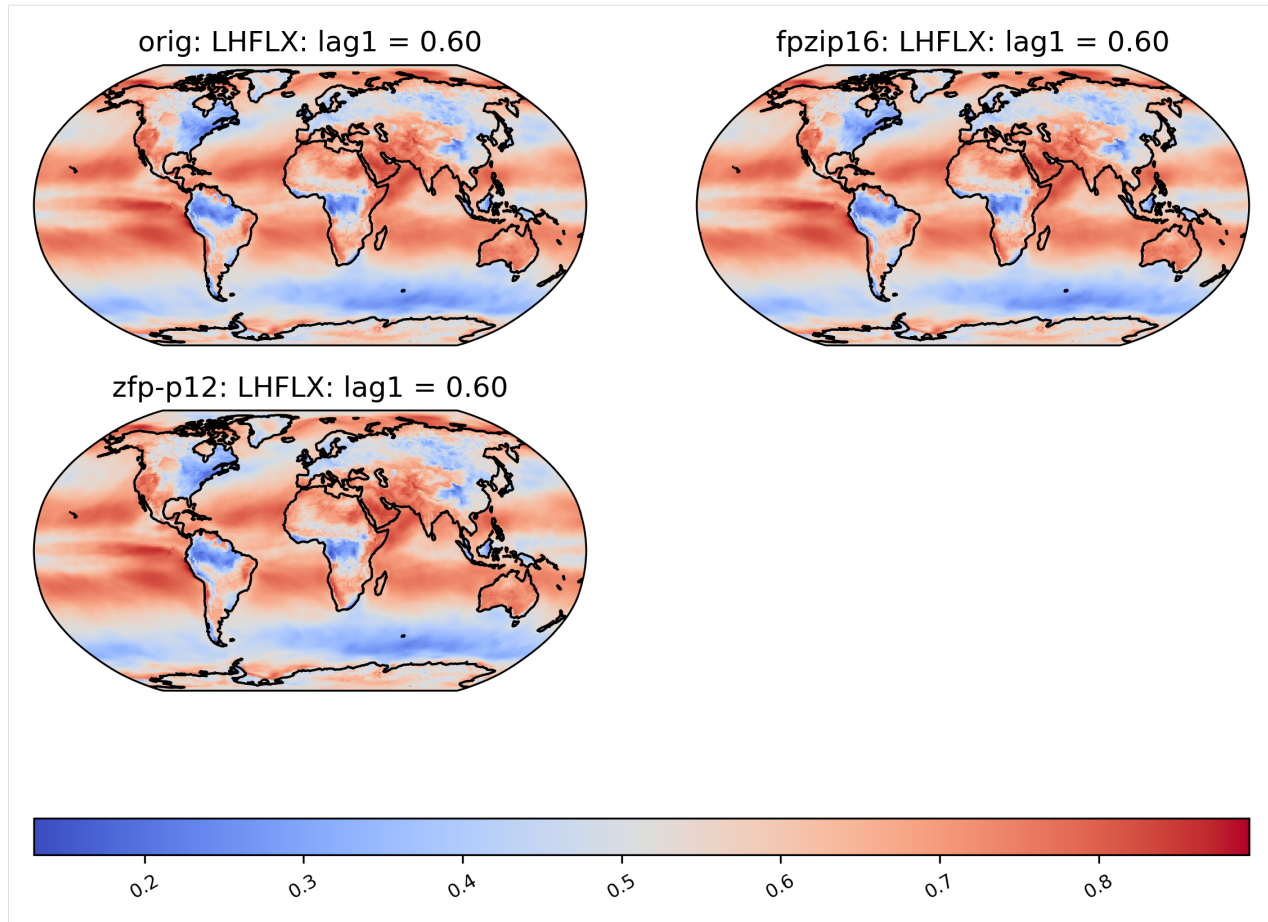
    LHFLX      (collection, time, lat, lon) float32 dask.array<chunksize=(1, 100, 192,
↳288), meta=np.ndarray>
Attributes: (12/14)
  Conventions: CF-1.0
  source:      CAM
  case:        b.e11.BRCP85C5CNBDRD.f09_g16.031
  title:       UNSET
  logname:     mickelso
  host:        ys1023
  ...
  initial_file: b.e11.B20TRC5CNBDRD.f09_g16.031.cam.i.2006-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:      Wed Nov 11 19:15:10 2020: ncks -L 5 LHFLX.daily.2006010...
  NCO:          netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures: area: cell_area
  data_type:    cam-fv

```

```

[18]: # plot of lag-1 correlation of LHFLX values for the first 10 years (NOTE: daily data
↳takes longer)
ldcpy.plot(
    col_lhflx,
    "LHFLX",
    sets=["orig", "fpzip16", "zfp-p12"],
    calc="lag1",
    start=0,
    end=3650,
)

```



```
[19]: # List the compressed data files for Q
os.listdir("/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q")
```

```
[19]: ['zfp.p18.Q.monthly.192001-200512.nc',
      'zfp.p10.Q.monthly.200601-208012.nc',
      'zfp.p16.Q.monthly.200601-208012.nc',
      'zfp.p20.Q.monthly.200601-208012.nc',
      'zfp.p8.Q.monthly.200601-208012.nc',
      'zfp.p10.Q.monthly.192001-200512.nc',
      'zfp.p22.Q.monthly.200601-208012.nc',
      'fpzip20.Q.monthly.192001-200512.nc',
      'fpzip20.Q.monthly.200601-208012.nc',
      'zfp.p18.Q.monthly.200601-208012.nc',
      'zfp.p12.Q.monthly.200601-208012.nc',
      'zfp.p20.Q.monthly.192001-200512.nc',
      'zfp.p22.Q.monthly.192001-200512.nc',
      'zfp.p14.Q.monthly.200601-208012.nc',
      'zfp.p8.Q.monthly.192001-200512.nc',
      'zfp.p14.Q.monthly.192001-200512.nc',
      'zfp.p12.Q.monthly.192001-200512.nc',
      'zfp.p16.Q.monthly.192001-200512.nc']
```

```
[20]: col_q = ldcpy.open_datasets(
```

(continues on next page)

(continued from previous page)

```

    "cam-fv",
    ["Q"],
    [
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/orig/Q.monthly.200601-
↪208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/fpzip20.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p8.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p10.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p12.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p14.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p16.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p18.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p20.Q.monthly.
↪200601-208012.nc",
        "/glade/p/cisl/asap/abaker/compression_samples/cam-lens/Q/zfp.p22.Q.monthly.
↪200601-208012.nc",
    ],
    [
        "orig",
        "fpzip20",
        "zfp-p8",
        "zfp-p10",
        "zfp-p12",
        "zfp-p14",
        "zfp-p16",
        "zfp-p18",
        "zfp-p20",
        "zfp-p22",
    ],
    chunks={"time": 100},
)
col_q

```

dataset size in GB 59.72

```

[20]: <xarray.Dataset>
Dimensions:      (collection: 10, time: 900, lev: 30, lat: 192, lon: 288)
Coordinates:
  * lat          (lat) float64 -90.0 -89.06 -88.12 -87.17 ... 88.12 89.06 90.0
  * lev          (lev) float64 3.643 7.595 14.36 24.61 ... 957.5 976.3 992.6
  * lon          (lon) float64 0.0 1.25 2.5 3.75 5.0 ... 355.0 356.2 357.5 358.8
  * time         (time) object 2006-02-01 00:00:00 ... 2081-01-01 00:00:00
    cell_area    (lat, collection, lon) float64 dask.array<chunksize=(192, 1, 288)>,
↪meta=np.ndarray>
  * collection   (collection) <U7 'orig' 'fpzip20' ... 'zfp-p20' 'zfp-p22'

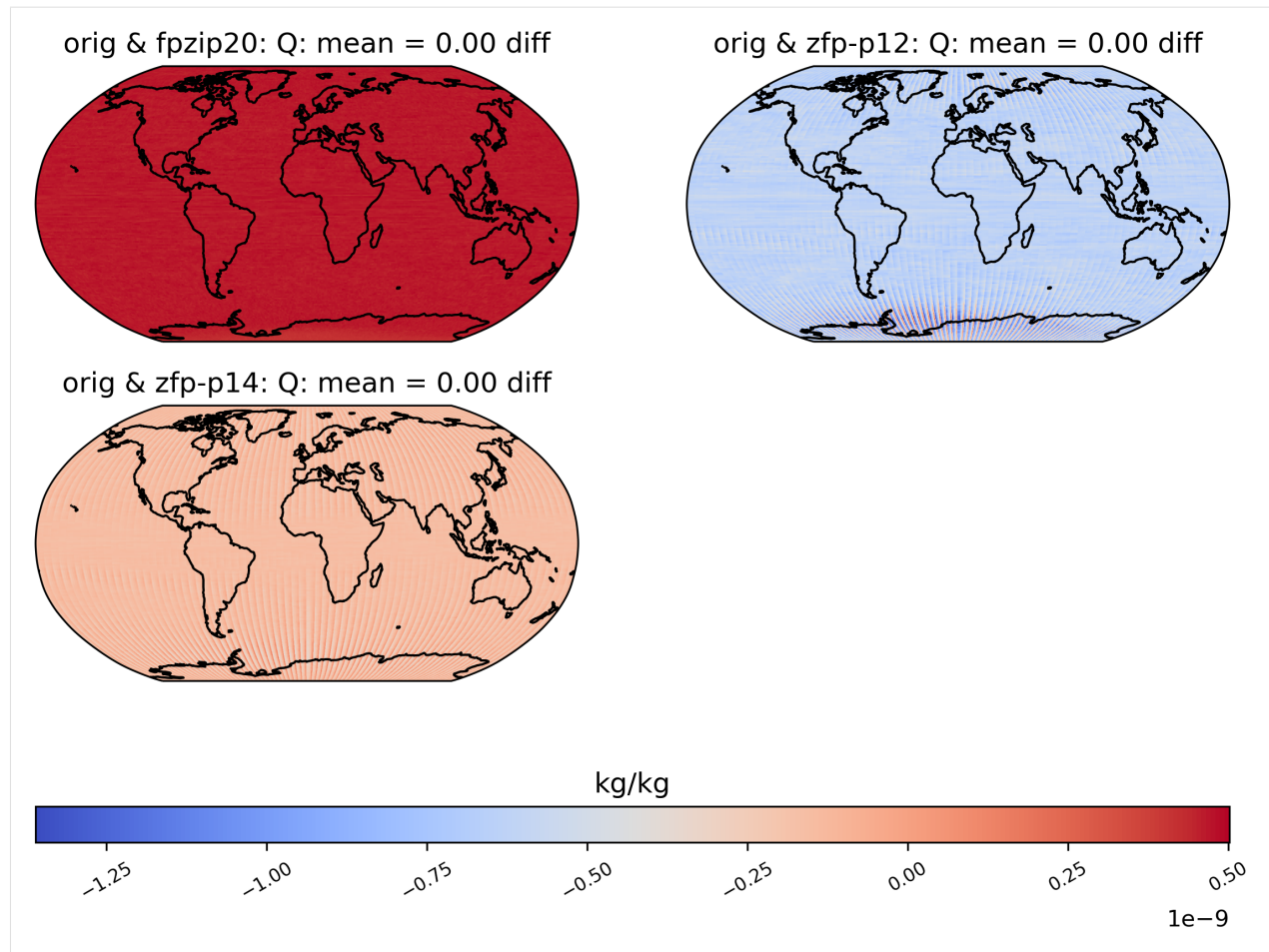
```

(continues on next page)

(continued from previous page)

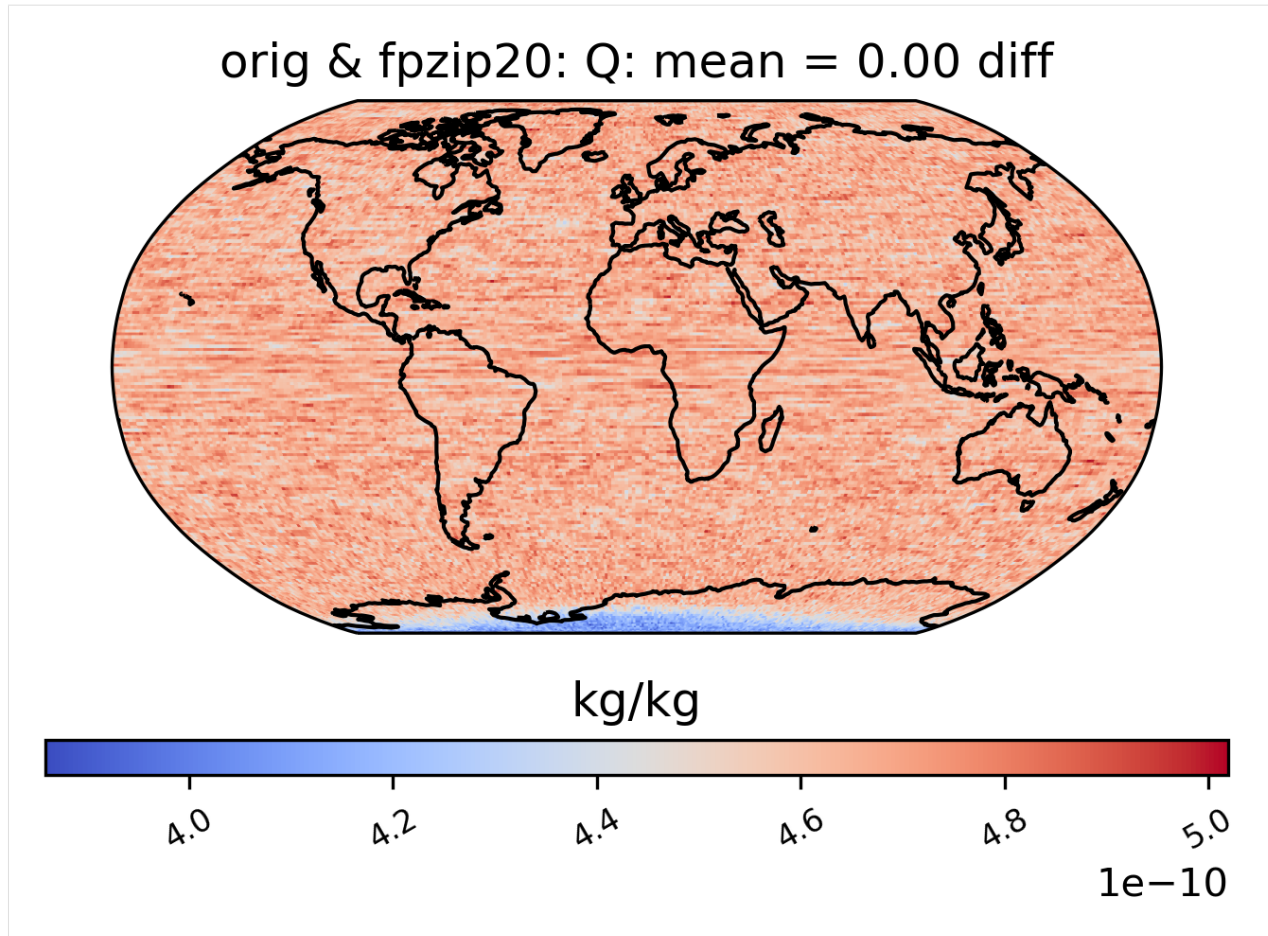
```
Data variables:
  Q      (collection, time, lev, lat, lon) float32 dask.array<chunks=(1, 100, 30, 192, 288), meta=np.ndarray>
Attributes: (12/14)
  Conventions: CF-1.0
  source:      CAM
  case:        b.e11.BRCP85C5CNBDRD.f09_g16.031
  title:       UNSET
  logname:     mickelso
  host:        ys1023
  ...
  initial_file: b.e11.B20TRC5CNBDRD.f09_g16.031.cam.i.2006-01-01-000000.nc
  topography_file: /glade/p/cesmdata/cseg/inputdata/atm/cam/topo/USGS-gtop...
  history:      Wed Nov 11 17:21:59 2020: ncks -L 5 Q.monthly.200601-20...
  NCO:          netCDF Operators version 4.7.9 (Homepage = http://nco.s...
  cell_measures: area: cell_area
  data_type:    cam-fv
```

```
[21]: # diff between mean Q values across the entire timeseries
ldcpy.plot(
    col_q, "Q", sets=["orig", "fpzip20", "zfp-p12", "zfp-p14"], calc="mean", calc_type=
    "diff", lev=0
)
```



```
[22]: # diff between mean Q values across the entire timeseries - look at just one from above
ldcpy.plot(col_q, "Q", sets=["orig", "fpzip20"], calc="mean", calc_type="diff", lev=0)
```

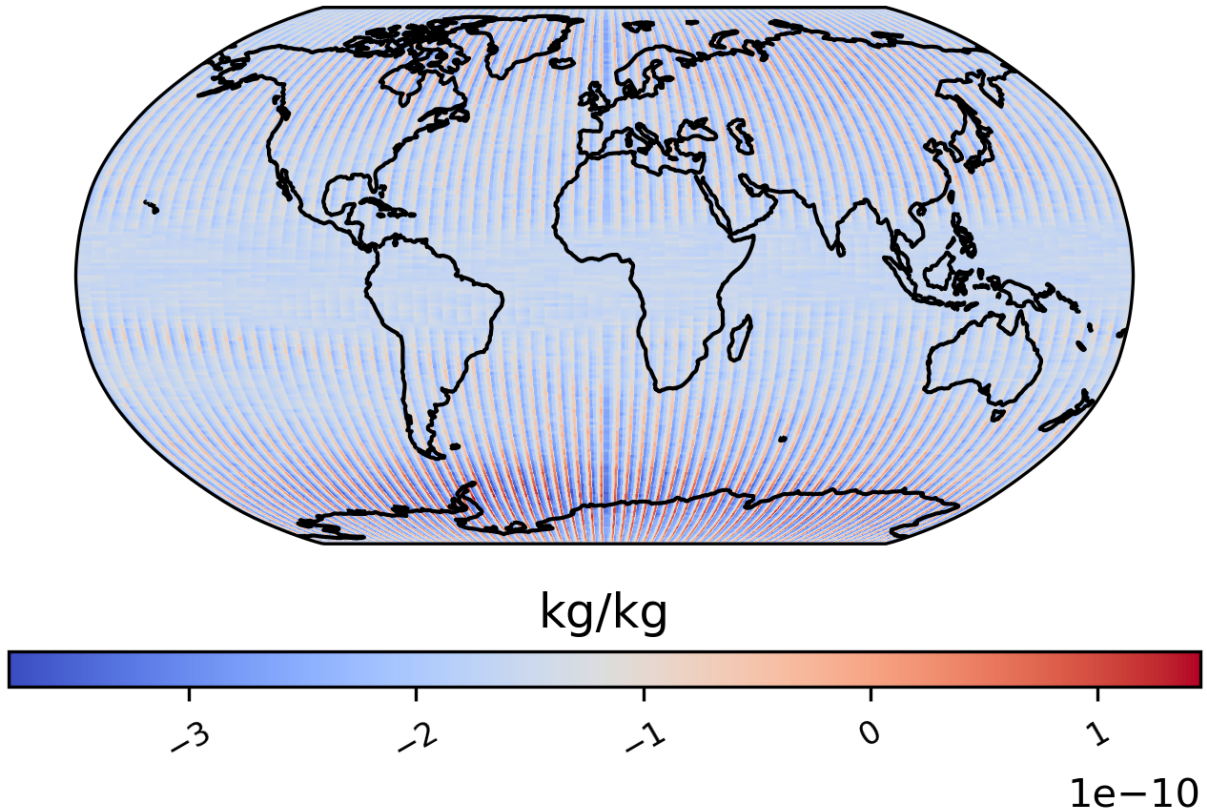




```
[23]: # diff between mean Q values across the entire timeseries - look at just one from above
ldcpy.plot(col_q, "Q", sets=["orig", "zfp-p14"], calc="mean", calc_type="diff", lev=0)
```



orig & zfp-p14: Q: mean = 0.00 diff



```
[24]: # Note: since q is 3D, need to select a level and a time slice
ldcpy.compare_stats(col_q.isel(time=0, lev=0), "Q", ["orig", "fpzip20", "zfp-p12", "zfp-
→p14"])
```

<IPython.core.display.HTML object>

	orig	fpzip20	zfp-p12	zfp-p14
mean	2.1926e-06	2.1921e-06	2.1932e-06	2.1927e-06
variance	9.558e-15	9.5586e-15	9.5603e-15	9.5582e-15
standard deviation	9.7765e-08	9.7768e-08	9.7777e-08	9.7766e-08
min value	2.1375e-06	2.1374e-06	2.1383e-06	2.1374e-06
max value	3.2387e-06	3.2382e-06	3.241e-06	3.2382e-06
probability positive	1	1	1	1
number of zeros	0	0	0	0
spatial autocorr - latitude	0.99169	0.99169	0.99156	0.99168
spatial autocorr - longitude	0.99917	0.99916	0.9991	0.99916
entropy estimate	0.34719	0.045749	0.021016	0.045467

<IPython.core.display.HTML object>

<IPython.core.display.HTML object>

	fpzip20	zfp-p12	zfp-p14
max abs diff	9.311e-10	6.7889e-09	2.0948e-09
min abs diff	0	0	0

(continues on next page)

(continued from previous page)

mean abs diff	4.6648e-10	1.1153e-09	3.0979e-10
mean squared diff	2.176e-19	3.6135e-19	2.4506e-20
root mean squared diff	5.3848e-10	1.366e-09	3.8902e-10
normalized root mean squared diff	0.0004859	0.001297	0.00037408
normalized max pointwise error	0.00084549	0.0057336	0.0015572
pearson correlation coefficient	1	0.99991	0.99999
ks p-value	6.4126e-18	1.1829e-51	6.837e-07
spatial relative error(% > 0.0001)	75.96	88.196	58.992
max spatial relative error	0.00043437	0.0026751	0.00074946
Data SSIM	0.80848	0.62469	0.84677

```
[25]: # Disconnect when finished
      cluster.close()
      client.close()
```

```
[ ]:
```

## INDICES AND TABLES

- `genindex`
- `modindex`
- `search`



## PYTHON MODULE INDEX

|  
`ldcpy.calcs`, [13](#)



## A

annual\_harmonic\_relative\_ratio  
     *(ldcpy.calcs.Datasetcalcs property)*, 14  
 annual\_harmonic\_relative\_ratio\_pct\_sig  
     *(ldcpy.calcs.Datasetcalcs property)*, 14

## C

calcsPlot (*class in ldcpy.plot*), 11  
 cdf (*ldcpy.calcs.Datasetcalcs property*), 14  
 check\_metrics() (*in module ldcpy.util*), 8  
 collect\_datasets() (*in module ldcpy.util*), 9  
 compare\_stats() (*in module ldcpy.util*), 9  
 covariance (*ldcpy.calcs.Diffcalcs property*), 16

## D

Datasetcalcs (*class in ldcpy.calcs*), 13  
 Diffcalcs (*class in ldcpy.calcs*), 16

## E

entropy (*ldcpy.calcs.Datasetcalcs property*), 14  
 ew\_con\_var (*ldcpy.calcs.Datasetcalcs property*), 14

## G

get\_calc() (*ldcpy.calcs.Datasetcalcs method*), 13  
 get\_diff\_calc() (*ldcpy.calcs.Diffcalcs method*), 16  
 get\_single\_calc() (*ldcpy.calcs.Datasetcalcs method*), 14

## K

ks\_p\_value (*ldcpy.calcs.Diffcalcs property*), 16

## L

lag1 (*ldcpy.calcs.Datasetcalcs property*), 14  
 lag1\_first\_difference (*ldcpy.calcs.Datasetcalcs property*), 14  
 lat\_autocorr (*ldcpy.calcs.Datasetcalcs property*), 14  
 ldcpy.calcs  
     module, 13  
 ldcpy.plot  
     module, 11  
 ldcpy.util

module, 8

(*ld-* lev\_autocorr (*ldcpy.calcs.Datasetcalcs property*), 14  
 lon\_autocorr (*ldcpy.calcs.Datasetcalcs property*), 14

## M

mae\_day\_max (*ldcpy.calcs.Datasetcalcs property*), 14  
 max\_spatial\_rel\_error (*ldcpy.calcs.Diffcalcs property*), 16  
 mean (*ldcpy.calcs.Datasetcalcs property*), 15  
 mean\_abs (*ldcpy.calcs.Datasetcalcs property*), 15  
 mean\_squared (*ldcpy.calcs.Datasetcalcs property*), 15  
 module  
     ldcpy.calcs, 13  
     ldcpy.plot, 11  
     ldcpy.util, 8  
 most\_repeated (*ldcpy.calcs.Datasetcalcs property*), 15  
 most\_repeated\_percent (*ldcpy.calcs.Datasetcalcs property*), 15

## N

n\_s\_first\_differences (*ldcpy.calcs.Datasetcalcs property*), 15  
 normalized\_max\_pointwise\_error (*ldcpy.calcs.Diffcalcs property*), 16  
 normalized\_root\_mean\_squared (*ldcpy.calcs.Diffcalcs property*), 16  
 ns\_con\_var (*ldcpy.calcs.Datasetcalcs property*), 15  
 num\_negative (*ldcpy.calcs.Datasetcalcs property*), 15  
 num\_positive (*ldcpy.calcs.Datasetcalcs property*), 15  
 num\_zero (*ldcpy.calcs.Datasetcalcs property*), 15

## O

odds\_positive (*ldcpy.calcs.Datasetcalcs property*), 15  
 open\_datasets() (*in module ldcpy.util*), 10

## P

pearson\_correlation\_coefficient (*ldcpy.calcs.Diffcalcs property*), 16  
 percent\_unique (*ldcpy.calcs.Datasetcalcs property*), 15  
 plot() (*in module ldcpy.plot*), 11

`pooled_variance` (*ldcpy.calcs.Datasetcalcs* property),  
15  
`pooled_variance_ratio` (*ldcpy.calcs.Datasetcalcs*  
property), 15  
`prob_negative` (*ldcpy.calcs.Datasetcalcs* property), 15  
`prob_positive` (*ldcpy.calcs.Datasetcalcs* property), 15

## R

`range` (*ldcpy.calcs.Datasetcalcs* property), 15  
`root_mean_squared` (*ldcpy.calcs.Datasetcalcs* prop-  
erty), 15

## S

`save_metrics()` (in module *ldcpy.util*), 10  
`spatial_rel_error` (*ldcpy.calcs.Diffcalcs* property),  
16  
`ssim_value` (*ldcpy.calcs.Diffcalcs* property), 16  
`ssim_value_fp_fast` (*ldcpy.calcs.Diffcalcs* property),  
16  
`ssim_value_fp_orig` (*ldcpy.calcs.Diffcalcs* property),  
16  
`ssim_value_fp_slow` (*ldcpy.calcs.Diffcalcs* property),  
17  
`standardized_mean` (*ldcpy.calcs.Datasetcalcs* prop-  
erty), 15  
`std` (*ldcpy.calcs.Datasetcalcs* property), 15  
`subset_data()` (in module *ldcpy.util*), 10

## T

`tex_escape()` (in module *ldcpy.plot*), 13  
`time_series_plot()` (*ldcpy.plot.calcsPlot* method), 11

## V

`variance` (*ldcpy.calcs.Datasetcalcs* property), 15

## W

`w_e_derivative` (*ldcpy.calcs.Datasetcalcs* property),  
16  
`w_e_first_differences` (*ldcpy.calcs.Datasetcalcs*  
property), 16

## Z

`zscore` (*ldcpy.calcs.Datasetcalcs* property), 16  
`zscore_cutoff` (*ldcpy.calcs.Datasetcalcs* property), 16  
`zscore_percent_significant` (*ld-  
cpy.calcs.Datasetcalcs* property), 16